

1-1-1979

# Psychometric and methodological contributions to criterion-referenced testing technology.

Daniel R. Eignor

*University of Massachusetts Amherst*

Follow this and additional works at: [https://scholarworks.umass.edu/dissertations\\_1](https://scholarworks.umass.edu/dissertations_1)

---

## Recommended Citation

Eignor, Daniel R., "Psychometric and methodological contributions to criterion-referenced testing technology." (1979). *Doctoral Dissertations 1896 - February 2014*. 3491.

[https://scholarworks.umass.edu/dissertations\\_1/3491](https://scholarworks.umass.edu/dissertations_1/3491)

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact [scholarworks@library.umass.edu](mailto:scholarworks@library.umass.edu).

UMASS/AMHERST



312066013538335

PSYCHOMETRIC AND METHODOLOGICAL CONTRIBUTIONS TO  
CRITERION-REFERENCED TESTING TECHNOLOGY

A Dissertation Presented

By

DANIEL R. EICNOR

Submitted to the Graduate School of the  
University of Massachusetts in partial fulfillment  
of the requirements for the degree of

DOCTOR OF EDUCATION

September

1979

EDUCATION

© Daniel R. Eignor 1979

All Rights Reserved

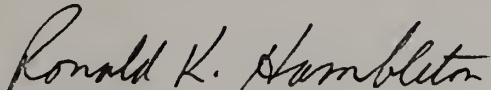
PSYCHOMETRIC AND METHODOLOGICAL CONTRIBUTIONS TO  
CRITERION-REFERENCED TESTING TECHNOLOGY

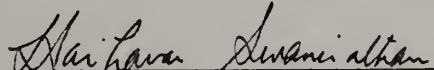
A Dissertation Presented

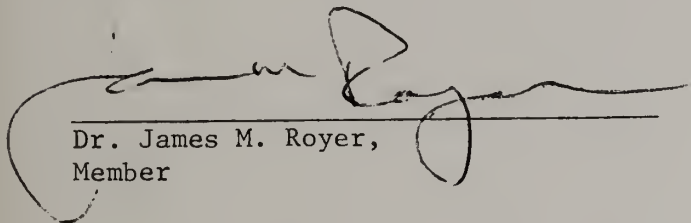
By

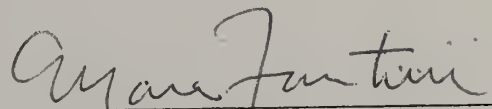
DANIEL R. EIGNOR

Approved as to style and content by:

  
\_\_\_\_\_  
Dr. Ronald K. Hambleton,  
Chairperson of Committee

  
\_\_\_\_\_  
Dr. Hariharan Swaminathan,  
Member

  
\_\_\_\_\_  
Dr. James M. Royer,  
Member

  
\_\_\_\_\_  
Dr. Mario Fantini, Dean  
School of Education

## A C K N O W L E D G E M E N T S

There are a number of individuals who were both instrumental and helpful in the educational experience I undertook for which this dissertation signifies the culmination. I would first like to thank my parents for the support they gave me. Never quite able to understand why I would want to leave a successful high school teaching career, they none-the-less were highly supportive of my venture, particularly during stressful personal moments I was to encounter.

I would be quite remiss if I did not recognize the initial impetus and subsequent help offered by my ex-wife Joan. Though we were to experience many personal problems during my graduate years, she always was supportive of my studies. Further, Joan was responsible for the initial encouragement that helped me get started in graduate school in the first place, for which I am most appreciative.

I would like next to acknowledge the help and assistance of my committee members, Swami and Mike, and my advisor, Ron. Mike, who made the initial contact that brought me to the University of Massachusetts, has always been most supportive of my coursework and the research presented in this dissertation. Further, my involvement with him in a number of Educational Psychology courses was a most positive and rewarding experience. Swami, in his initial personal support and subsequent academic support both in his classes and on this dissertation, was also most instrumental in my achieving this degree. Finally,

I am hardpressed to express the gratitude I have for Ron's assistance. He quite simply has allowed me the opportunity to work closely with him in a highly supportive fashion. The end result has been, besides this dissertation, a fine understanding of the field of criterion-referenced testing and the opportunity to partake in a number of rewarding professional experiences. In sum, I view the guidance and advice afforded me by my committee members and advisor as being of a most positive nature, and I am grateful for it.

I'd like to recognize next a number of friends and fellow students who were instrumental in my achieving my degree. My good friends Merilee, Harry, and Greta Neunder should be thanked for the support and just plain fun they gave me during my graduate years. I'd like to thank Janice Gifford both for the continued friendship shown me and for the assistance she gave me with my dissertation research. Rick DeFriesse was also most helpful in getting the simulation program running that was used for the research reported in Chapter 3. I'd like to thank Anne Fitzpatrick for adding her certain kind of humor and levity to my final years of study. Bernie McDonald should be recognized for her persistence in typing this dissertation, but more importantly, for the friendship shown me. I'd be remiss if I did not recognize my supervisors at ETS, Al Carlson and Jane Houis.. Their assistance in helping me complete my dissertation during a difficult first year of employment is most appreciated. I'd also like to thank good friends Louise Fox and Kent Poey for the help afforded me during the months surrounding my divorce. They were most instrumental in keeping me going personally, and hence academically. Finally, I would like to thank my close and dear friend Linda Cook for everything she has done

for me over the past three years. These have been difficult personal years for me, and Linda has been most instrumental in keeping me positively directed.



ABSTRACT

Psychometric and Methodological Contributions to  
Criterion-Referenced Testing Technology

(September 1979)

Daniel R. Eignor, B.S., Manhattan College  
M.S., SUNY at Albany  
Ed.D., University of Massachusetts, Amherst

Directed By: Ronald K. Hambleton

The launching of Sputnik in the late 1950's, and the ensuing educational emphasis placed on individualized instruction, was to foster impetus for a new testing technology designed to answer the questions of what it is an individual student does or does not know. This testing technology, called criterion-referenced testing, has known a somewhat uneven start in reference to formalization of methods of test development, assessment of psychometric properties, and on a more rudimentary level, simple definitions of terms. The last time anyone bothered to count, there were over 600 references on the topic of criterion-referenced testing. Unfortunately, it seems that there have been almost as many ideas about what a criterion-referenced test is as there have been contributions to the field. Recently, however, a number of researchers in the field have published integrating works that have improved the situation greatly. Definitional problems have been resolved and a criterion-referenced test development process has been articulated.

Much work in the field still remains to be done. A survey of the field demonstrates that one of the most pressing problems for measurement specialists of today has been the necessity to produce criterion-referenced test technology and instruments quickly. Unfortunately, the desire of many individuals, organizations, and agencies to use criterion-referenced tests has far outrun the testing profession's ability to produce test development standards and high quality instruments to meet this need. As a consequence, classroom teachers have been constructing "home-made" tests or using commercially prepared criterion-referenced tests of undetermined quality to make instructional decisions; program evaluators, recognizing the shortcomings of norm-referenced tests in program evaluation activities have been constructing criterion-referenced tests based on the best psychometric principles they can find in a body of literature that is confusing, contradicting, and substantially gap-laden; and professional licensing organizations have been grappling with issues such as the determination of cut-off scores in the midst of complicated legal actions. The three situations described above, as well as many others, have contributed to an unsettled situation in the field of criterion-referenced testing. The three problem areas discussed above form the basis for the research reported in this dissertation.

The first problem area addressed in this dissertation is the present lack of a suitable set of guidelines for the development and evaluation of criterion-referenced tests and test manuals. Most of the major test publishers have published in the last few

years a wide assortment of criterion-referenced tests. In addition, many school districts, state agencies, small testing firms, and consulting firms have produced their own criterion-referenced instruments. However, a review of these instruments will demonstrate that the majority of tests fall short of the technical quality necessary for them to accomplish their intended purposes. A reasonable explanation for this situation is that there has been a shortage of usable guidelines for constructing and using criterion-referenced tests. In this dissertation, a set of 39 guidelines are offered with a rationale and procedures for applying them to the evaluation of criterion-referenced tests. These guidelines are then applied to eleven of the more popular commercially prepared criterion-referenced tests in the field to demonstrate that the proposed guidelines are workable and to highlight areas where considerably more work on the part of test developers is needed.

The second problem addressed in this dissertation involves a psychometric area of criterion-referenced testing research that is essentially unexamined, the relationship of test length to reliability and validity. A primary concern of all individuals using test scores is that the scores be both valid and reliable. While it is well-known that there is a direct relationship between the length of a test and the reliability and validity of the test scores, little has been done in the field of criterion-referenced testing to articulate the relationship. In this dissertation, the relationship is investigated via simulation techniques and tables

relating test length to reliability and validity indices are offered for a wide variety of situations. These tables hopefully will help practitioners in the field make some practical decisions about suitable criterion-referenced test lengths.

The third area investigated in this dissertation is perhaps the most critical due to the present emphasis placed in the nation's schools on minimum competency testing. While there presently exist a variety of well-known methods for setting cut-off scores, there does not exist a suitable set of guidelines to help the concerned individual select a method from the myriad of cut-score methods. Proper selection of a method is important because existing methods are based on differing assumptions. In this dissertation, the myriad of methods suitable for criterion-referenced standard setting are first organized into a number of categories and then applied to a prototypical testing situation, minimum competency testing. Recommendations about methods for use are offered and one of these methods, the Modified Angoff Technique, is applied to minimum competency tests in the Insurance field. Implementation strategies are offered to aid practitioners interested in applying this method in their work.

## TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS. . . . .	iv
ABSTRACT. . . . .	vii
LIST OF TABLES. . . . .	xiv
LIST OF FIGURES . . . . .	xvii
 CHAPTER	
I INTRODUCTION. . . . .	1
1.1 Background. . . . .	1
1.2 Statement of the Problems . . . . .	9
1.3 Purposes. . . . .	12
1.4 Organization of the Study . . . . .	13
II GUIDELINES FOR EVALUATING CRITERION-REFERENCED TESTS AND TEST MANUALS. . . . .	14
2.1 Introduction. . . . .	14
2.2 Review of the Literature. . . . .	16
2.3 Methods of Investigation. . . . .	21
2.3.1 Development of the Guidelines	
2.3.2 Guidelines	
2.3.3 Development of the Evaluation Form	
2.3.4 Choice of Tests for Evaluation	
2.3.5 Application of the Guidelines to the Tests	
2.4 Results and Discussion. . . . .	56
2.5 Conclusion. . . . .	66
III THE RELATIONSHIP OF TEST LENGTH TO CRITERION- REFERENCED TEST RELIABILITY AND VALIDITY. . . . .	68
3.1 Introduction. . . . .	68
3.2 Some Background Information . . . . .	70
3.2.1 Norm-Referenced Approaches to Reliability and Validity	
3.2.2 Two Criterion-Referenced Test Score Uses	

3.3	Reliability and Validity for Mastery State Assignments. . . . .	74
3.3.1	Reliability	
3.3.2	Validity	
3.4	Criterion-Referenced Test Length. . . . .	88
3.4.1	Hsu's Study of Test Length- Reliability	
3.5	Research Methodology. . . . .	98
3.5.1	Variables Under Investigation	
3.5.2	Simulation Procedures	
3.6	Results and Discussion. . . . .	109
3.7	Conclusions . . . . .	134
IV	SETTING STANDARDS FOR CRITERION-REFERENCED TESTS AND AN APPLICATION TO MINIMUM COMPETENCY TESTING. . . . .	137
4.1	Introduction. . . . .	137
4.2	Methods for Setting Cut-Off Scores Suitable for Criterion-Referenced Tests. . . . .	140
4.2.1	Judgmental Models—Item Content	
4.2.2	Judgmental Models—Guessing and Item Sampling	
4.2.3	Empirical Models—Data From Two Groups	
4.2.4	Decision-Theoretic Procedures	
4.2.5	Empirical Models Depending Upon a Criterion Measure	
4.2.6	Educational Consequences	
4.2.7	Combination Models: Judgmental- Empirical	
4.2.8	Combination Models: Bayesian Procedures	
4.3	Setting Cut-Scores for Minimum Competency Tests. . . . .	184
4.4	An Application of the Method Angoff Procedure. . . . .	193
4.4.1	Background Information	
4.4.2	Choice of a Method for Determining a Cut-Score	
4.4.3	Panel System Design	
4.4.4	Panel Tasks—Question Rating Form	
4.4.5	Panel Tasks—Content Rating Form	
4.4.6	Results—Cut-Off Scores	
4.4.7	Results—Content Validity	
4.4.8	Comments on Cut-Score Procedures	
4.5	Conclusions . . . . .	220

CHAPTER	Page
V CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH. . . . .	222
REFERENCES. . . . .	226
APPENDIX. . . . .	234

LIST OF TABLES

Table		Page
2.3.1	Criterion-Referenced Tests Reviewed in the Study. . . . .	50
2.3.2	Summary of the Criterion-Referenced Tests. . . . .	52
2.4.1	Number and Percentage of Tests Reviewed in Each Rating Category on the Guidelines for Evaluating Criterion-Referenced Tests and Test Manuals . . . . .	58
3.6.1	Simulated Distributions Considered in the Study. . . . .	110
3.6.2	Selected Test Lengths and Associated Cut-Offs for Simulation One: Equal and Unequal Weights . . .	115
3.6.3	Selected Test Lengths and Associated Cut-Offs for Simulation Two: Equal and Unequal Weights . . .	117
3.6.4	Selected Test Lengths and Associated Cut-Offs for Simulation Three: Equal Weights . . . . .	118
3.6.5	Selected Test Lengths and Associated Cut-Offs for Simulation Four: Equal Weights. . . . .	119
3.6.6	Selected Test Lengths and Associated Cut-Offs for Simulation Five: Equal and Unequal Weights. . .	122
3.6.7	Selected Test Lengths and Associated Cut-Offs for Simulation Six: Equal and Unequal Weights. . .	123
3.6.8	Selected Test Lengths and Associated Cut-Offs for Simulation Seven: Equal Classification Weights. . . . .	125
3.6.9	Selected Test Lengths and Associated Cut-Offs for Simulation Eight: Equal Weights . . . . .	127
3.6.10	Selected Test Lengths and Associated Cut-Offs for Simulation Nine: Equal Weights. . . . .	129



Table	Page
3.6.11 Selected Test Lengths and Associated Cut-Offs for Simulation Ten: Equal Weights. . . . .	130
3.6.12 A Comparison of Subkoviak (1) and Subkoviak (2) Estimates of Proportion Agreement . . . . .	133
3.6.13 The Relationship of Cut-Off Scores to the Simulated Distribution for Three Selected Distributions . . . . .	135
4.3.1 A Comparison of Several Standard Setting Methods . . . . .	190
4.4.2 Estimates of Average Number of Answers Known by the Minimally Knowledgeable Applicant Group (Question Rating Form). . . . .	207
4.4.3 Number of Questions a Minimally Competent Person Would Answer Correctly (Content Rating Form)—Life Test . . . . .	208
4.4.4 Number of Questions a Minimally Competent Person Would Answer Correctly (Content Rating Form)—Accident and Health Test. . . . .	209
4.4.5 Number of Questions a Minimally Competent Person Would Answer Correctly (Content Rating Form)—Property Test . . . . .	210
4.4.6 Number of Questions a Minimally Competent Person Would Answer Correctly (Content Rating Form)—Casualty Test . . . . .	211
4.4.7 A Comparison of the Cut-Off Scores for the Two Procedures Used—Life Test. . . . .	212
4.4.8 Number of Questions Judged Content Appropriate by 75% or Greater of the Judges—Life Test. . . . .	215
4.4.9 Number of Questions Judged Content Appropriate by 75% or Greater of the Judges—Accident and Health Test . . . . .	215
4.4.10 Number of Questions Judged Content Appropriate by 75% or Greater of the Judges—Property Test. . . . .	216
4.4.10 Number of Questions Judged Content Appropriate by 75% or Greater of the Judges—Casualty Test. . . . .	216

Table	Page
4.4.12 Panel Responses Regarding the Appropriateness of Content Area—Life Test. . . . .	217
4.4.13 Panel Responses Regarding the Appropriateness of Content Area—Accident and Health Test . . . . .	217
4.4.14 Panel Responses Regarding the Appropriateness of Content Area—Property Test. . . . .	218
4.4.15 Panel Responses Regarding the Appropriateness of Content Area—Casualty Test. . . . .	218

LIST OF FIGURES

Figure	Page
3.6.1 Percentage distributions for simulations one thru four. . . . .	114
3.6.2 Quartiles and difficulty and ability parameters for simulations five thru ten . . . . .	121
4.2.1 Three general sets of procedures for setting standards. . . . .	141
4.2.2 A classification of models and methods for determining standards. . . . .	148
4.3.1 A classification of models and methods for determining minimum competency standards . . . . .	189
4.4.1 Pictorial representation of panel formulation and sequential order of tasks. . . . .	201
4.4.2 Instructions for Question Rating Form . . . . .	203
4.4.3 Instructions for Content Rating Form . . . . .	206

# CHAPTER I

## INTRODUCTION

### 1.1 Background

The reawakening of the field of *criterion-referenced measurement* can be linked to the launching of Sputnik in the late 1950's.<sup>1</sup> In an attempt to interpret why the Soviet Union's space program was ahead of ours, the group-based approach to education in the United States came under close scrutiny. Sputnik was interpreted as a sign that our educational system was not keeping pace with the Soviet's, and from this interpretation, was to come a demand for accountability and an interest in alternative educational approaches. While this new focus on accountability and innovation caused many changes in curriculum and instruction, it also gave evidence to the fact that the traditional, or *norm-referenced*, testing and measurement practices, which had been perfected to a high level of sophistication, were no longer useful for the new testing situations encountered. A new testing and measurement methodology was necessary to focus on the accountability issue and to measure the effects of the new instructional procedures. The established norm-referenced testing methodology focused on the construction of tests that facilitated the comparison of individual examinees. Such comparative data is not useful for addressing the accountability issue,

---

<sup>1</sup>The birth of criterion-referenced measurement came in the 1920's and early 1930's with the growth of interest in individualized instruction (Washburne, 1922). Block (1971) has provided an excellent review of this earlier movement.

where program evaluation is the concern (Popham, 1978a) or for making decisions about what an individual student does or does not know (Hambleton, Swaminathan, Algina, and Coulson, 1978).

Before discussing the new criterion-referenced testing methodology however, it is worthwhile to present further information on both the innovative approaches developed and on the accountability movement.

Traditional approaches to instruction were primarily group-based. It was assumed that pupils entered school with aptitudes and skills that spread them out along the normal curve. When they graduated, the students were still spread out along the normal curve. The educational experience brought about a shift of the distribution along the proficiency continuum, but the experience produced few shifts in how students were distributed (Popham, 1978a). With Sputnik, such a viewpoint came under close inspection. Instead of focusing on groups of pupils and the normal curve, the focus fell instead on the individual student. Group-based instructional approaches and the relative comparison of students gave way to concern about instruction focused on the individual and the assessment of what the individual student did or did not know. The idea was to present learning activities that helped each individual optimize his or her potential. The idea of a fixed position on an ability distribution was held in disregard (see Carroll, 1963; Bloom, 1968). The development of teaching machines and programmed instruction signaled the advent of this change of focus to the individual, and with this change came the clear realization that traditional or norm-referenced measurement wouldn't work. Such tests did not give a clear indication of what an examinee

could or could not do, and this information was critical for planning individual learning opportunities. Thus, with the development of programs stressing the individual, it became clear that a new measurement methodology was necessary. Glaser (1963), while in the process of examining programmed instructional techniques, was perhaps the first individual to clearly describe this need for new measurement practices, and he called the needed new measurement techniques *criterion-referenced measurement*.

While the launching of Sputnik was to help precipitate a switch in emphasis to the education of the individual, the publication of the Project Talent data in 1964 (Flanagan, Davis, Dailey, Shaycoft, Orr, Goldberg, & Neyman, 1964) clearly documented the need for significant change in elementary and secondary schools. This need for change brought about the development of a number of educational programs that stressed the individualization of instruction in an attempt to improve the educational experience (Gibbons, 1970; Gronlund, 1974; Heathers, 1972). These programs, which are somewhat related to the earlier programmed instruction movement, all have a common characteristic; curricula are defined by of instructional objectives. Examples of such programs include Individually Prescribed Instruction (Glaser, 1968, 1970), Program for Learning in Accordance With Needs (Flanagan, 1967, 1969) and Mastery Learning (Carroll, 1963, 1970; Bloom, 1968; Block, 1971). Hambleton (1974) has provided a comprehensive review of these instructional programs. All of the programs share as their goal the provision of an educational program that is maximally adaptive to the requirements of the individual.

The instructional objectives specify the curriculum and serve as the basis for the development of curriculum materials and achievement tests. According to Hambleton, Swaminathan, Algina, and Coulson (1975):

One of the underlying premises of objectives-based programs is that effective instruction depends, in part, on a knowledge of what skills the student has. It follows that the tests used to monitor student progress should be closely matched to instruction. [p. 2]

Thus, it can be seen that a measuring instrument to be used in assessing student performance should be keyed to instruction and also provide information that can be used to make decisions on an individual basis. Further, a measuring instrument should provide information that can be used to measure student progress along an absolute achievement continuum. Given these stipulations, it is once again evident why norm-referenced measuring instruments are of limited use for these programs. Such instruments or tests are constructed to facilitate the making of comparisons across students, and hence are not well suited for making the sorts of decisions required by individualized instructional programs. Stated in a different fashion, norm-referenced tests "are principally designed to produce test scores suitable for ranking individuals on the ability measured by the test" (Hambleton & Novick, 1973). When the question is whether or not an individual has achieved a specific level of mastery, a comparison of the student to other examinees will not answer the question. The basic purposes of testing in

these individualized programs have been expressed by Hambleton and Novick (1973):

It would seem that in most cases, the pertinent question is whether or not the individual has attained some prescribed degree of competence on an instructional performance task. Questions of precise achievement levels and comparisons among individuals on these levels seem to be largely irrelevant. In many of the new instructional models, tests are used to determine on which instructional objectives an examinee has met the acceptable performance level standard set by the model designer. This test information is usually used immediately to evaluate the student's mastery of the instructional objectives covered on the test, so as to appropriately locate him for his next instruction.  
[p. 160]

Thus far, in this introduction, the change in instructional emphasis brought about by the events following the launching of Sputnik has been discussed. Little has been said about the accountability movement, which was generated out of concern that schools might not be doing their job. Without going into great detail about the logistics of the accountability movement, one relevant comment can be made. For program evaluators and administrators involved in the assessment of program effects, it quickly became evident that norm-referenced test score data was not going to answer the accountability question. Popham (1978a) and Hambleton and Gifford (1977) have discussed the limitations of norm-referenced test data for program evaluation. Three reasons offered by both Popham and by Hambleton and Gifford for these limitations are:

1. There is seldom congruence between the content covered by the norm-referenced test and the content of the program being evaluated. This is because norm-referenced tests are based on an amalgamation of



objectives of traditional programs in various sections of the country, and hence, are not truly representative of any one program.

2. Norm-referenced tests do not provide the information necessary to improve poorly functioning programs. Further, the tests are usually built on the objectives of traditional programs, and thus are not suitable for assessing innovative programs.
3. Because norm-referenced tests are constructed to spread students out along a continuum, items that contribute to test score variability are selected. Therefore, items tapping concepts that are taught successfully by a majority of teachers will be removed from the test. The test then becomes an instrument sensitive to the aptitude of the student's rather than to the effects of instruction. If the test is to be a measure of the instructional process, the content should be matched to the program. Norm-referenced test construction techniques, in maximizing test score variability, can destroy the match between test content and the program's objectives.

In conclusion, the events following Sputnik ushered in an emphasis on individualized instructional programs in conjunction with an interest in the thorough evaluation of the effects of such programs. Further, the measurement problems inherent in these programs and the public demand for accountability necessitated the development of a new, criterion-referenced testing methodology.

As mentioned earlier in this chapter, Glaser (1963) was the first individual to introduce and define criterion-referenced measurement. However, after Glaser's initial work little of a developmental nature in the field was done until the important Popham and Husek paper was published in 1969. Since 1969, research in this area has increased at what seems to be an exponential rate; at present, there are over 600 references on criterion-referenced testing (Hambleton et al., 1978). On the application level, there are at

present millions of students at all levels of education taking criterion-referenced tests. (These tests are also referred to as *performance-based*, *skills-oriented*, or *competency-based*.) For example, criterion-referenced tests are being used to monitor individual progress through objectives-based instructional programs, to diagnose learning deficiencies, to evaluate educational and social action programs, and to assess examinee competencies on various certification and licensing examinations.

Glaser and Nitko (1971) offered one of the most popular definitions of a criterion-referenced test: "A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards." Performance standards do *not* refer to normative performance levels, but rather "The performance standards are usually specified by defining some domain of tasks that the student should perform. Representative samples of tasks from this domain are organized into a test. Measurements are taken and are used to make a statement about the performance of each individual relative to that domain."

Popham (1975) has provided a more recent definition of a criterion-referenced test that parallels Glaser and Nitko's: "A criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavior domain." If one utilizes either Glaser and Nitko's or Popham's definition, the construction of a criterion-referenced test requires sampling of items from well-specified domains (of items). The domain "may be

extensive or a single, narrow objective, but it must be well defined, which means that content and format limits must be well specified" (Millman, 1974). The specification of the domain is crucial for putting together a criterion-referenced test since only then can the scores be most directly interpreted in terms of knowledge of performance tasks (Hambleton et al., 1978). Popham's definition is similar to the definition offered by Millman (1974) for a *domain-referenced test*, and so, if Popham's definition is adopted, the two descriptions (criterion-referenced test and domain-referenced test) may be used interchangeably. However, criterion-referenced tests are different from *objectives-referenced tests*, which are constructed to match behavioral objectives. According to Hambleton et al. (1978):

The primary distinction between criterion-referenced tests and objectives-referenced tests is as follows: In a criterion-referenced test, the items are a representative set of items from a clearly defined domain of behavior measuring an objective, whereas with an objectives-referenced test no domain of behavior is specified, and items are not considered to be representative of any behavior domain. [p. 3]

In conjunction with a number of papers that present a variety of definitions of criterion-referenced measurement, many papers reflecting diverse views on methods of test development, the assessment of psychometric properties, and criterion-referenced test applications have been written. However, with the integrating work of Hambleton and Novick (1973), Harris, Alkin, and Popham (1974), Millman (1974), Popham (1975, 1978a), Hambleton et al. (1978), and Hambleton and Eignor (1978a), the situation has greatly improved.

There now exists sufficient theory and guidelines for implementing criterion-referenced testing programs as far ranging as objectives-based instructional programs at the classroom level, program evaluations at the district and statewide level, and competency-based certification programs at the state and national level. Further, guidelines for criterion-referenced test construction and validation relevant for the practitioner are now available (Hambleton & Eignor, 1978a).

While the literature is presently in a more coherent state than it was in the early 1970's, a number of problems in this area remain to be solved. In the next section, the specific problems addressed in this research study will be introduced.<sup>2</sup>

## 1.2 Statement of the Problems

From a careful review of the present state of criterion-referenced testing technology, three problem areas were identified that required resolution in order to ensure that criterion-referenced tests could serve their intended purposes. While many other problem areas have been identified by Hambleton et al. (1978), Popham (1978a), and Hambleton and Eignor (1978a), the three selected for study seem especially important.

The first problem area in criterion-referenced testing is the present lack of a suitable set of usable guidelines for the development and evaluation of criterion-referenced tests and test manuals.

---

<sup>2</sup>The research reported in this dissertation was supported by a Basic Skills Research Grant awarded to Ronald K. Hambleton by the National Institute of Education in the summer of 1978.

Most of the major test publishers (Harcourt - Brace - Jovanovich, CTB/McGraw-Hill, Houghton-Mifflin, and Science Research Associates) have published in the last few years a wide assortment of criterion-referenced tests. In addition, many school districts, state agencies, small testing firms, and consulting firms have produced their own criterion-referenced tests. However, from a review of these available instruments, and from discussions with others who have reviewed the tests, it is evident that most of the tests fall short of the technical quality necessary for them to accomplish their intended purposes. A reasonable explanation for this situation is that there has been a shortage of usable guidelines for constructing and using criterion-referenced tests. The well-known Standards for evaluating tests and test manuals prepared by the joint committee of AERA/APA/NCME is helpful, but those Standards are not completely applicable to criterion-referenced tests. The other research done in this area (Popham, 1978a; Walker, 1977; Swezey & Pearlstein, 1975) represents a start in the right direction, but usable sets of guidelines were not produced. Hambleton and Eignor (1978a, 1978b), in some initial work, offered a set of guidelines and an overall evaluation of many currently available criterion-referenced tests. However, they provided no rationale for their choice of guidelines or detailed analysis of the tests they reviewed.

A second problem area in criterion-referenced testing concerns the relationship of criterion-referenced test length to test score reliability and validity. A primary concern of all individuals using test scores is that the test scores be both valid and reliable.

While the best approaches to assessing reliability and validity are situation-specific, it is well-known that there is a direct relationship between the length of a test and the reliability and validity of the test scores. Longer tests result in test scores with better psychometric properties.

While the present criterion-referenced testing literature abounds with research papers on the subjects of reliability (Livingston, 1972a, 1972b, 1972c; Swaminathan, Hambleton, & Algina, 1974; Hambleton & Novick, 1973; Huynh, 1976; Subkoviak, 1976), validity (Cronbach, 1971; Messick, 1975; Linn, 1977; Livingston, 1978) and test length (Millman, 1973; Novick & Lewis, 1974; Phaner, 1974; Wilcox, 1976), there is only one paper that this author is aware of that investigated the relationship of test length and reliability. This is an unpublished paper by Hsu (1977), and because of the underlying model involved in Hsu's formulation, his results, are not very useful. The only work done to date relating criterion-referenced test length to validity has been an unpublished paper by Livingston (1978), and in this paper, that relationship is only indirectly investigated.

The third and final problem area has to do with the problem of cut-off scores to be used for assignment of individuals to mastery states. While there exist a variety of methods for the setting of cut-off scores, and most of these methods are well-known (Millman, 1973; Meskauskas, 1976; Hambleton & Eignor, 1978a), there at present does not exist a suitable set of guidelines to help the concerned individual select a method from the myriad of cut-off score methods.

Proper selection of a method is important because existing methods are based on different sets of assumptions, and the assumptions underlying the method chosen should be appropriate for the situation in which the method will be used. Further, once a method for setting cut-off scores has been decided upon, implementation strategies usually do not exist to aid in the actual use of the chosen method. A report by Livingston and Zieky (1977) presents one of the few procedures for practitioners to follow to obtain cut-off scores. Thus, individuals concerned about the setting of cut-off scores, be it for mastery learning situations or more importantly for this dissertation, minimum competency examinations, have little guidance available at present to aid in the selection and implementation of a method for setting a cut-off score.

### 1.3 Purposes

The purposes of this study were linked directly to the three problems discussed in the last section. In reference to the first problem discussed, the lack of usable guidelines for evaluating criterion-referenced tests, the following objectives guided the research:

1. Development of a set of usable guidelines (with appropriate rationale offered for their inclusion) for use in the evaluation of criterion-referenced tests and test manuals.
2. Application of the guidelines to the evaluation of several standardized criterion-referenced tests, and the preparation of a report of the results.

Research in this area was intended to serve as a follow-up to earlier work initiated by Hambleton and Eignor (1978a, 1978b).

In reference to the second problem area, the relationship of test length to criterion-referenced reliability and validity, the

following objectives guided the research:

1. Development of a computer program to relate test lengths to criterion-referenced reliability and validity indices.
2. Completion of a simulation study relating ability distributions, test lengths, cut-off scores, domain score estimates, test score characteristics, and loss ratios, to a variety of reliability and validity indices.
3. Preparation of a set of tables relating test length to reliability and validity under a wide variety of testing conditions.

The research on cut-off scores, which comprised the last area of investigation, was guided by the following objectives:

1. Organization of the available methods for setting cut-off scores in a useful form for practitioners.
2. Presentation of guidelines and implementation strategies to aid individuals in answering the following questions: "How can the 'best' method for use in a prototypical situation be selected?" and "How should the chosen method be implemented?". The prototypical situation selected for presentation in this dissertation involved minimum competency testing.

#### 1.4 Organization of the Study

The remainder of the study is organized around four chapters. Chapter II is devoted to the Guidelines developed for the evaluation of criterion-referenced tests and test manuals. Chapter III contains the research that relates criterion-referenced test lengths to reliability and validity, and Chapter IV contains the work done on the organization of available methods for setting cut-off scores, in conjunction with the selection and implementation strategies developed. Finally, Chapter V contains suggestions for further research in the three areas of criterion-referenced testing investigated in this dissertation.



C H A P T E R   I I  
GUIDELINES FOR EVALUATING CRITERION-REFERENCED  
TESTS AND TEST MANUALS

2.1 Introduction

Most of the major test publishers have published in the last few years a wide assortment of criterion-referenced tests. In addition, many school districts, state agencies, small testing firms, and consulting firms have produced their own criterion-referenced tests. These tests are designed to address many measurement areas. For example, criterion-referenced tests are being used to monitor student progress through school programs, to diagnose learning disabilities, to report student progress to parents, to evaluate various types of programs, and to certify or license professionals in many fields. Unfortunately, many of the available tests fall short of the technical quality necessary for them to accomplish their intended purposes (Hambleton & Eignor, 1978b). One explanation for this observation is that many criterion-referenced tests were developed before an adequate testing technology was fully explicated. Fortunately, there now exists an adequate technology for constructing criterion-referenced tests and using criterion-referenced test scores (Hambleton & Eignor, 1978a; Hambleton, Swaminathan, Algina & Coulson, 1978; Popham, 1978a). Another explanation for this observation is that there has been a

shortage of guidelines for constructing and using criterion-referenced tests. Certainly the well-known Standards for evaluating tests and test manuals prepared by a joint committee of AERA/APA/NCME is helpful, but it is not completely applicable to criterion-referenced tests. Besides the incompleteness of the AERA/APA/NCME Standards for evaluating criterion-referenced tests and test manuals, what relevant information there is, is scattered through 75 pages or so of other materials appropriate for norm-referenced test evaluations. Therefore, the Standards in its present form, is not very useful for individuals interested in evaluating criterion-referenced tests.

A review of the criterion-referenced testing literature resulted in the location of three articles or books that attempted to develop guidelines for evaluating tests and test manuals. These include a report by Swezey and Pearlstein (1975), an unpublished manuscript by Walker (1977), and Chapter 8 of Popham's book, Criterion-Referenced Measurement (1978a). While these guidelines will be reviewed in greater detail in the next section of this chapter, at this point it should be stated that this author feels that these guidelines either lack completeness or are of highly subjective nature, and hence, are of somewhat limited use for evaluating a wide variety of criterion-referenced tests and test manuals.

Hambleton and Eignor (1978a, 1978b) offered a set of guidelines and an overall evaluation of many currently available criterion-referenced tests. However, they provided no rationale for their choice of guidelines or detailed analysis of the tests they studied.

The primary purpose of the research presented here was to expand upon the initial set of guidelines developed by Hambleton and Eignor. Guidelines are offered along with a rationale for inclusion and a set of ratings to be used with each guideline. The guidelines are offered as a set of questions for consideration by potential users and developers of criterion-referenced tests.

The research reported in this chapter was also guided by another purpose. Hambleton and Eignor (1978b), in applying the guidelines to eleven standardized criterion-referenced tests, were able to offer only a general overall evaluation. In this chapter, more specific details of the applications of the guidelines to the tests are offered.

In summary, the research reported in this chapter was guided by the following two objectives:

1. Development of a set of guidelines (with appropriate rationale offered for their inclusion) for use in the evaluation of criterion-referenced tests and test manuals.
2. Application of the guidelines to the evaluation of eleven selected standardized criterion-referenced tests, and a preparation of a complete report of the results.

## 2.2 Review of the Literature

The available literature relating to guidelines for evaluating criterion-referenced tests and test manuals is meager. The only three efforts to develop such guidelines that this author is aware of are contained in Swezey and Pearlstein's (1975) Guidebook for Developing Criterion-Referenced Tests, prepared for the U.S. Army

Research Institute for the Behavioral and Social Sciences; Walker's (1977) manuscript on Standards and the related CSE Test Evaluation books (particularly CSE Secondary School Test Evaluations); and Popham's (1978a) eighth chapter in his book on Criterion-Referenced Measurement. Before discussing these materials in greater depth, two other sources that were of help in developing the guidelines presented in this chapter should be cited and briefly discussed. These are the Standards for Educational and Psychological Tests (1974) prepared by a joint committee of APA, AERA, and NCME, and the chapter on selecting and evaluating tests in F. G. Brown's (1976) excellent text Principles of Educational and Psychological Testing (2nd ed.).

A reading of the Standards published by the APA/AERA/NCME Committee leaves one impressed with the comprehensive coverage given the various aspects of test design and interpretation. The 73 pages of Standards cover a wide range of concerns that are listed under the following general headings: dissemination of information, aids to interpretation, directions for administration and scoring, norms and scales, validity, reliability and measurement error, qualifications and concerns of users, choice or development of test or method, administration and scoring, and interpretation of scores. There are, however, certain difficulties involved with these Standards; these difficulties being both of a general nature, and also, specifically in terms of application to criterion-referenced tests. On a general level, there are two problems that can be mentioned. One, because the Standards are so all-encompassing, they are unwieldy. This

unwieldiness is also related to the second problem; there is no suitable evaluation form that an individual can work with in evaluating a test. On the specific level of evaluating criterion-referenced tests, the following observation can be made. The Standards are really applicable to norm-referenced tests, and what material there is that is relevant to criterion-referenced tests is spread throughout the Standards booklet. An individual would need to go through a "sifting process" to apply the Standards in a reasonable fashion to a criterion-referenced test.

Brown (1976) discusses procedures for selecting and evaluating standardized tests. Of particular importance, Brown has taken the Standards developed by APA/AERA/NCME and worked them into a ten category format for test evaluations. This evaluation form is most useful for the evaluation of norm-referenced tests, thereby alleviating one of the major problems with the Standards. The form, in the state Brown has presented it, is not directly applicable to criterion-referenced tests. The form was, however, most useful as a starting point in developing the guidelines reported in this chapter.

Swezey and Pearlstein's (1975) Guidebook for Developing Criterion-Referenced Tests was at its time of publication, perhaps the most comprehensive statement of procedures for developing criterion-referenced tests. Popham's (1978a) book, and the work of Hambleton and Eignor (1978a), now offer the test developer a number of other sources on criterion-referenced test development to choose from. Swezey and Peralstein's Guidebook was prepared for use by the Army, and hence the focus is slanted toward military testing procedures.

There is, however, a short discussion on guidelines that was consulted in the development of the guidelines presented in this chapter.

Walker's (1977) paper and the related CSE MEAN Evaluation System (reported in CSE Secondary School Test Evaluations) represents another attempt to develop guidelines. A perusal of the MEAN Evaluation System leaves the reader both impressed and at the same time, somewhat concerned. The System is certainly comprehensive in coverage, having 39 criteria upon which to evaluate a test. However, the concern on the part of the reader comes from an arbitrary weighting system of the criteria. Further, weights on each criterion are summed (for four categories) to give an overall rating. No information is provided on how the weights were chosen; further, since weights differ across criteria, information should be presented on why one criterion is weighted more heavily than another. As an example, the following two criteria relating to the test's "Administrative Usability" are presented verbatim from the text.

- a. To how large a group can the test be administered? For purposes of classroom or school evaluation it is important to economize on the time and effort in the administration of tests. If the test can be administered to groups of more than 35, according to the recommendations of the test manual, the test was credited with 2 points; if the group must number less than 35, the test was credited with 1 point; and if the test must be administered on an individual basis, the test was credited with 0 points.
- b. Is the norm group representative of the national population? Six considerations comprised the evaluation of the representatives of the groups used to norm the test: (1) Was the sample obtained through cluster, stratified, or random, rather than incidental sampling? (2) Was the norming done less than 5 years ago? (3) Was there geographic representation? (4) Was the appropriate age range represented and

exhausted? (5) Was there racial/ethnic representation or were separate norms available? (6) Were population density characteristics (e.g., urban, suburban, rural, etc.) represented? If the answer to these questions, based upon convincing tabulations for the third, fourth, fifth, and sixth over, was "yes" for five or six of them, the text was credited with 1 point. If there were fewer than five "yes" answers, the test was credited with 0 points.

This author does not dispute the importance of both criteria, but rather, questions the points allotted to the two, and would have preferred to see an explanation of the point allocation system included. However, in fairness to the CSE group, the MEAN Evaluation System was developed for their own use in evaluating tests, and was not offered as a general evaluation system for public use.

Popham (1978a) presented six characteristics of a well-constructed criterion-referenced test. These characteristics should be sought in the evaluation of criterion-referenced tests, as their absence limits the usability of the test in question. These six characteristics are

1. Unambiguous description—in the manual, the test developer has to describe exactly what a test score is an indication of. According to Popham; "But lengthy or terse, the critical quality of these descriptive schemes is that they permit one to make an unequivocal description of what a test taker's performance truly signifies."
2. Sufficient items—there must be an adequate number of items to measure each behavior that is being tested.
3. Appropriate focus—the test must measure a manageable and interpretable number of behaviors.
4. Reliability—the test must consistently measure the defined behavior.
5. Validity—the objectives and items must be valid. Further, the test must serve the purpose for which it was constructed.

6. Comparative data—the test manual should provide data on how other examinees perform on the test (i.e., normative data).

Popham's discussion of these six characteristics is on a general level; he does not offer an evaluative instrument in conjunction with them. His discussion was, however, useful to the development of the guidelines suggested in this chapter.

In summary, the guidelines by Swezey and Pearlstein (1975), Walker (1977) and the CSE Evaluation Group (1974), and Popham (1978a) were used as background material for the guidelines offered here.

## 2.3 Methods of Investigation

### 2.3.1 Development of the Guidelines

In this section, a brief description of the procedures involved in the development of the guidelines is presented. The first step in the development was a thorough review of the Standards offered by APA/AERA/NCME. Those standards that were applicable to criterion-referenced tests were removed and placed in the appropriate ten categories of Brown's (1976) suggested format for Test Evaluations. It was found that in placing the APA/AERA/NCME Standards into Brown's categories, it was necessary to delete the General Information category and add categories on Manual Preparation, Qualifications of Test Users, and Test Interpretations.

The second step in the development of the guidelines involved a merger of the material prepared in step one, the guidelines and/or



suggestions offered by Swezey and Pearlstein (1975), Walker (1977), and Popham (1978a), material included in the criterion-referenced review article by Hambleton et al. (1978) and the instructional materials prepared by Hambleton and Eignor (1978a). Non-relevant material generated through step one was removed from the guidelines, and material reflecting the suggestions of the other authors listed above, and recent advances in the field, was added.

The third step in developing the materials involved the preparation of a list of guidelines. This was obtained by placing ourselves (Hambleton & Eignor, 1978b) in the role of potential purchasers of a criterion-referenced test, and asking "What questions would we want to answer before making a decision to use a criterion-referenced test in a particular situation?" The questions generated were organized around ten broad categories, which include Objectives, Test Items, Administration, Test Layout, Reliability, Cut-off Scores, Validity, Norms, Reporting of Test Score Information, and Test Score Interpretations.

Finally, in the last step involved in the development of the guidelines, a rationale for the inclusion of each guideline was prepared, along with a rating scale for judging tests vis à vis each guideline.

### 2.3.2 Guidelines

In this section, the guidelines for evaluating criterion-referenced tests and test manuals are presented. With each guideline is included a rationale for inclusion and ratings for evaluating a test or test manual.

Objectives:

- A.1. Is the purpose (or purposes) of the test stated in a clear and concise fashion?

Rationale:

It is very important that the purpose or purposes for the test be stated in a clear, concise fashion (preferably) in the introductory section of the test manual. Such information will aid a potential user in making a decision about whether the test satisfies his/her needs.

Ratings:

Acceptable: The purpose is stated in a understandable fashion in one particular section or paragraph of the introduction.

Acceptable with reservations: The purpose is stated in the introductory section, but is fragmented, such that information must be drawn from various paragraphs.

Unacceptable: There is no clear statement of the purpose for the test. The potential user then must decide on test purpose and from that, whether he/she wants to use the test.

- A.2. Is each objective clearly written so that it is possible to identify an "item pool"?

Rationale:

The identification of an "item pool" is very important so as to increase the likelihood of valid inferences about examinee performance. The test user (usually) wants to make an inference, based upon test scores, about an examinee's level of performance in the "domain" of behaviors being measured. In order to do this, the domain must be well-defined so that test items can be viewed as a random sample from the domain.

Ratings:

Acceptable: Each objective is written so that appropriate content and difficulty is clear. There should be no possibility that potential users will differ in their understanding of relevant item pools.

Unacceptable: An "item pool" can't be identified. For example, item writers would differ significantly in the content they use to write items matched to the same objective.

A.3. Is it clear from the list of objectives what the test measures?

Rationale:

From a list of behavioral objectives, the test user can get an idea of what the test is measuring, and probably make a decision about whether or not the test is suitable. Thus, at this level of decision-making, the use of behavioral objectives would be sufficient. However, when inferences are to be made, based only upon specifications of objectives, there are problems. Tests developed from a specification of objectives are called "objectives-referenced tests," and it should be understood that the best interpretation that a test user can make about examinee performance will be test-specific. Valid inferences can't occur.

Ratings:

Acceptable: The content of the test is specified through the use of behavioral objectives. The test user can easily make a subjective decision about what the test measures.

Unacceptable: The content of the test is not clearly defined so that the potential user can't make a decision about what is being measured.

A.4. Is an appropriate rationale offered for including each objective in the test?

Rationale:

The test manual should explain to the potential test user in clear terms why each objective in the test was included. Explanations could take the form of a statement of the importance of the objective in the content area, or the fact that content specialists agreed that the objective should be included. Regardless of form, there should be a statement to the user telling him/her why the objective was included.

Ratings:

Acceptable: There is a clear statement, either for each objective, or for all objectives considered together, as to why they were included on the test.

Acceptable with reservations: There is at least a general explanation about why objectives were included on the test.

Unacceptable: There is no statement as to why objectives were included, thereby forcing the potential user to make a subjective decision about the objectives.

- A.5. Can a user "tailor" the test to meet local needs by selecting objectives from a pool of available objectives?

Rationale:

Since criterion-referenced tests are used to decide what an individual examinee does or doesn't know in reference to a content area, the potential user should have the flexibility of selecting from the overall test those objectives that he/she feels should be administered to an individual. For instance, some school districts often want to select objectives to match their curriculum. Further, if the user is sure that an examinee has mastered certain objectives, then there is no reason to test for them. Thus, flexibility should be a part of tests being used to make individual diagnostic decisions.

Ratings:

Acceptable: The test user can select those objectives he/she wants to test. This feature will not always be of interest to potential users.

Acceptable with reservations: The test user can choose to administer sub-tests, made up of a number of objectives of which the objective of interest is a member. There is some, though limited, flexibility for the user.

Unacceptable: The test user must administer the whole test in an intact fashion. There is no way of subdividing the test into objectives; i.e., there is no information in the manual about how to do this.

- A.6. Is there a match between the content measured by the test and the situation where the test is to be used?

Rationale:

Since these criterion-referenced tests are used (for the most part) to make individual diagnostic decisions, there must be a suitable match between the objectives and the test-use situation for the diagnosis to take place. If the objectives test something other than what the test user is interested in, the test is of no use.

Ratings:

Acceptable: The objectives must be specific to a well-developed content area and specified clearly enough to provide information about what the examinee does or doesn't know, i.e., to be of diagnostic value.

Unacceptable: The objectives are not specific to a defined content area and thus will not provide valuable diagnostic information about what an examinee knows in a content area.

- A.7. Are individuals identified who were responsible for the preparation of objectives?

Rationale:

The test manual should identify who participated in the objective selection process. Further, it is not really enough to specify "specialists"; the manual should further identify area of specialty, such as "reading specialist"; "test specialist"; etc. To be most complete and informative, specialists who participated in the objective selection process should be identified by name, and their areas of specialty.

Ratings:

Acceptable: Those individuals who participated in the objective selection process are identified by name, and by area of specialty.

Acceptable with reservation: The individuals are identified only generally as, for instance, "specialists in the field of reading."

Unacceptable: There is no data supplied on who participated in the objective selection process and/or development process.

- A.8. Does the set of objectives measured by the test serve as a representative set from some content domain of interest?

Rationale:

The test manual should provide some information on how complete the set of objectives is in reference to the area or sub-areas of content being measured. This could, for instance, take the form of the identification of the views of the content experts involved about how complete the objectives set is.

Ratings:

Acceptable: The manual provides some data substantiating the state of the set of objectives concerning completeness of coverage of the subject area.

Acceptable with reservations: The manual provides no information on how complete the set of objectives is, but it does provide a comprehensive list of objectives, thereby allowing the test user some ease in deciding for him/herself about completeness of coverage.

Unacceptable: The manual affords no way, either objectively or subjectively, for the test user to decide on the completeness of the set of objectives.

B. Test Items

B.1. Is the item review process described?

Rationale:

In keeping with current methods for constructing criterion-referenced tests, a panel of content specialists should review the test items with two concerns in mind: (1) Are the domain specifications (or objectives if it is an objectives-based test) clearly written?, and (2) Do the items measure the domain (or objective)? Reporting of item analysis information is not sufficient to be considered an item review.

Ratings:

Acceptable: The item review process is clearly defined in the manual. The potential user thereby gains some assurance that the items do measure the content area.

Acceptable with reservations: The item review process is described in a general way; nothing is said about particular procedures utilized. The test user is able to determine that a review process occurred, but not how.

Unacceptable: Either nothing is said in the manual about how items were reviewed or it is identified that the items were reviewed based solely upon their statistical properties.

- B.2. Are the test items valid indicators of the objectives they were developed to measure?

Rationale:

If the item review process (B.1.) is done properly, the test user can be reasonably confident that the items are valid indicators of the objectives they were written to measure. Otherwise, the user must make a subjective decision about the items.

Ratings:

Acceptable: The item review process is clearly described in the manual, thereby assuring the user that the test items are valid.

Acceptable with reservations: The item review process is not clearly described, but there is a comprehensive list of objectives and an ample set of items so that the test user can convince him/herself that the items are valid.

Unacceptable: There are not sufficient items or a comprehensive list of objectives that would allow the user, even in a subjective fashion, to determine whether the items are valid indicators of the objectives.

- B.3. Is the set of test items measuring an objective representative of the "pool" of items measuring the objective?

Rationale:

In the selection of items for a criterion-referenced test, it is important that the items be a representative sample of the pool of items that could be generated to test the objective. Further, if the items are selected for inclusion in the sample based solely on statistical properties, the representative nature of the sample may be destroyed.

Ratings:

Acceptable: The items were not selected based upon statistical characteristics, and sufficient data is offered in the test manual to allow the test user to make an objective (or subjective) decision about whether the items are a representative sample.

Acceptable with reservations: Statistical data has been used in conjunction with judgmental data in selecting items, and the manual doesn't clearly sort out for the user which was more important in the decision process.

Unacceptable: Either no method is provided for the user to make a subjective decision (i.e., the objectives aren't listed) or the items were chosen based solely on statistical characteristics.

B.4. Are the items technically correct?

Rationale:

Proper item writing procedures should be followed in the construction of items for the tests. If not spoken of specifically in the manual, there should be sufficient data (items) for the user to subjectively convince him/herself of the fact.

Ratings:

Acceptable: The manual either informs the users that suitable procedures have been followed, or supplies sufficient information to allow the user to make a confident subjective decision.

Unacceptable: Proper item writing techniques have not been followed, as is evident from perusing the items.

B.5. Was a suitable format for the items selected?

Rationale:

A suitable format should be utilized for the items that are selected to be on the test. For instance, if the test is being used for diagnostic purposes, then the item format should be one that provides the most possible diagnostic information.

Ratings:

Acceptable: The item format fits the purpose for which the test is designed.

Unacceptable: An item format not congruent with the purpose of the test was chosen, thereby reducing the amount of information available.



- B.6. Are the test items free of bias (for example, sex, ethnic, or racial)?

Rationale:

If bias of any sort (sex, ethnic, racial, socioeconomic) is likely to be a concern for the content area being measured by the test, the test manual should identify any procedures utilized to remove the bias. (This is more likely to be a concern with language-based tests than with mathematical tests.)

Ratings:

Acceptable: The manual has identified procedures for removing item bias in those situations where it is likely to be a problem.

Unacceptable: Item bias has not been considered as a potential problem in relevant content areas.

- B.7. Was a heterogeneous sample of examinees employed in piloting the test items?

Rationale:

In keeping with current criterion-referenced testing technology, besides subjecting the items to a rigorous review process, they should also be piloted. Data can be collected to help view concerns such as the adequacy of the directions, etc., but also, the data collected on the items can be used to see which items aren't "working" properly and why. The test user should be informed as to the nature of the sample of examinees used in the pilot study.

Ratings:

Acceptable: The manual clearly describes the nature of the sample used in the pilot study.

Unacceptable: Any of the following situations has occurred: (1) a pilot study was run, but with a restricted sample, (2) no pilot study was run, or (3) the manual presents no data on this question.

- B.8. Was the item analysis data used only to detect "flawed" items?

Rationale:

Item analysis data, whether it be traditional norm-referenced indices such as item difficulty and item discrimination, or specialized CRT indices such as item sensitivity or Popham's chi-square, should be used to detect flawed items that require

rewriting and not as a criterion for item selection. Further, item analysis data should be collected and used before the final test is administered. That is, the focus in the use of these indices should be on detection of problems, and not demonstration that the test is measuring properly the objectives. A proper review process will assure that the items (provided they are properly constructed) are measuring the objectives.

Ratings:

Acceptable: The item analysis data was collected in the pilot or field-test stage and used to detect flawed items.

Unacceptable: The item analysis data was used as the criterion for inclusion of the items in the test at the test construction stage, or the item analysis data was collected after the test was constructed as a demonstration that the test is measuring properly the objectives.

C. Administration

C.1. Do the test directions include information relative to test purpose, time limits, practice questions, answer sheets, and scoring?

Rationale:

Information such as test purpose, time limits, practice questions, etc. should be contained both in the teacher's manual and in the explicit directions that the examinee reads or is read to him/her. The test taker should be informed of these issues prior to taking the exam. Also, the test administrator should have a clear idea of how to deal with these issues prior to administering the test. For instance, the test instructor should know how to deal with a question about whether or not a student should guess on an item.

Ratings:

Acceptable: Both the directions to the student and to the test administrator contain explicit statements of test purpose, time limits, practice questions, directions on how to use answer sheets, etc.

Unacceptable: There is no explicit statement of test purpose, etc., in either the directions to the test administrator or the individual examinee.

C.2. Are the test directions clear?

Rationale:

The test directions should be clearly written in both of the parts into which the directions can be separated, the information that the test administrator reads to him/herself and the part he/she reads to or has the examinees read. Otherwise, there will be problems when the actual test is administered.

Ratings:

Acceptable: Both the directions to the test administrator and to the examinee are clearly written, and the language of the directions is at a grade-level that can be comprehended by the student.

Unacceptable: The directions are poorly written to the extent that the test administrator is forced to interpret them in his/her own words.

C.3. Is the test easy to score?

Rationale:

If a test is difficult to score, errors generated through scoring mistakes will enter into the test scores. Further, those individuals who score the tests (be it the teacher or the student) will have difficulty and view the task negatively. Also, if the test is being used for diagnostic purposes, the errors generated from scoring difficulties may cause an improper diagnosis (i.e., the student is a master, but is diagnosed as a non-master). Of course, this is less of a problem if the test is machine scored.

Ratings:

Acceptable: The test should cause no difficulties in the scoring process. If the teacher is the person to score the test, it should be a simple task for him/her. If the manual says the students may score, then the task should be suitable for their grade level.

Unacceptable: The test is so difficult to score that scoring error enters substantially into the final (objective) scores obtained.

C.4. Does the test manual specify an examiner's role and responsibilities?

Rationale:

The examiner's role and responsibilities (what is expected of him/her during the test administration) should be clearly specified in the test directions. This will insure that the subsequent test administration will run smoothly.

Ratings:

Acceptable: The test directions clearly specify the examiner's role and responsibilities.

Unacceptable: Role and responsibilities of the test administrator are not clearly specified. Then, the examiner is forced to "ad-lib" procedures to facilitate test administration.

D. Test Layout

D.1. Is the layout of the test booklet attractive?

Rationale:

The layout of the test booklet should be attractive to the test taker. This will tend to minimize negative feelings, boredom, etc., and for the younger test-taker, surely generate some enthusiasm. The test should be fun for that age group.

Ratings:

Acceptable: The layout of the test booklet is attractive to the user.

Unacceptable: The layout of the test booklet is not attractive to the test taker.

D.2. Is the layout of the test booklets convenient for examinees?

Rationale:

The layout of the test booklet should be convenient for the examinee thereby minimizing frustration and confusion. For instance, if more than one objective is included per page of the booklet, the reading task becomes more difficult.

Ratings:

Acceptable: The layout of the test booklet causes no problems for examinees.

Unacceptable: There are problems in test layout that will cause the test-taking experience to be less than optimal.

## E. Reliability

E.1. Is the type of reliability information offered in the test manual appropriate for the intended use (or uses) of the scores?

### Rationale:

There are two primary uses (on the individual level) of test scores: estimation of domain scores (the score the individual would have obtained had he/she taken all the questions), and for instructional decision-making (allocating an individual mastery or non-mastery status). The reliability evidence should be consistent with the intended use of the scores. If domain score estimation is the intended use, an indication of the precision of the estimate should be offered. This can take the form of the standard error of measurement or the standard error of estimation derived from the binomial test model. The more precise the estimate (the smaller the error), the more reliable the test score is as an estimate of domain score.

If the test scores are being used to make instructional decisions, some indication of the consistency of decision-making over parallel forms or a retest administration should be offered. This could take the form of coefficient kappa, or a proportion of agreement index.

### Ratings:

Acceptable: The manual provides reliability evidence consistent with the intended uses of the criterion-referenced test.

Acceptable with reservations: The manual provides reliability evidence generated from an ad-hoc procedure similar to suitable procedures, or the manual provides some proper reliability evidence in conjunction with procedures that may not suitably demonstrate criterion-referenced test reliability.

Unacceptable: The manual provides reliability data generated from norm-referenced procedures, or from ad-hoc procedures that do not provide consistency evidence that coincides with the test score usage.

- E.2. Was the sample (or samples) of examinees used in the reliability study adequate in size, and representative of the population for whom the test is intended?

Rationale:

In order for the reliability evidence to be generalizable, the sample used must be large and representative of an appropriate population. Also, the size of the sample and the population from which the sample was drawn should be identified in the manual. The potential user can then check the applicability of the reliability information to his/her testing situation.

Ratings:

Acceptable: The manual clearly presents the size of the sample and a description of the population from which it was drawn.

Acceptable with reservations: Either the size of the sample is questionable or the degree of representativeness of an appropriate population is in question. The potential user is unable to clearly ascertain whether the reliability information is going to be applicable to his/her situation.

Unacceptable: The size of the sample and the degree of representativeness are inadequate, or no information is supplied in the manual.

- E.3. Are test lengths suitable to produce tests with desirable levels of test score reliability?

Rationale:

If the criterion-referenced test score is being used to estimate a domain score, precision of estimation is the critical indicant to be observed (see E.1.). There must be a sufficient number of items on the test for the test score to be a reasonable estimate of the domain score.

If the criterion-referenced test score is being used to make mastery/non-mastery decisions, it is critical that there be a sufficient number of test items to provide data to make the decision. Otherwise an unacceptably large number of false-positive and false-negative errors will occur. For instance, a student might guess correctly on one question, and on that basis alone be allotted mastery, when he/she really was not (a false-positive error). As another example, a student could incorrectly mark the answer to one question measuring an objective, and on

that basis alone be allotted non-mastery when he/she was in reality a master (a false-negative error). There needs to be a sufficient number of questions measuring each objective to minimize the role of errors in determining mastery status.

Ratings:

Acceptable: The number of items included to measure each objective is large enough for the test to be reliable, and the information about why that number was chosen is included in the manual.

Acceptable with reservations: The number of items appears large enough to the user, but inadequate information is offered in the manual as to how the number was decided upon.

Unacceptable: There is an insufficient number of items for each objective to be sure that the test will measure the objective in a reliable fashion (for the intended use).

- E.4. Is reliability information offered in the test manual for each intended use (or uses) of the test scores?

Rationale:

If the test scores produced are being used to make more than one sort of decision, reliability information should be offered for each use. For instance, in certain testing programs, a two-step testing procedure is used. First, a general test measuring a number of objectives is taken, and then, based upon the results, a number of mini-tests focusing on each objective may be taken. The point to be made here is that if it has been shown that the mini-tests are reliable, it can't be assumed that the general test is, although the tests measure basically the same content. They serve different uses; the first is used as an initial screen, the second for indepth diagnosis. Reliability evidence must be provided for each use.

Ratings:

Acceptable: The manual clearly provides reliability evidence for each of the intended uses of the test scores.

Acceptable with reservations: The manual provides reliability evidence for each of the uses, but in differing degrees of completeness. Certain of the usage areas have received inadequate investigation for establishing reliability.

Unacceptable: The manual provides little or no evidence of the reliability of the score(s).

#### F. Cut-Off Scores

- F.1. Was a rationale offered for the selection of a method for determining cut-off scores?

##### Rationale:

Besides using a suitable procedure for setting cut-off scores, the rationale behind the selection of the procedure should also be offered. This rationale should contain a discussion of the general underlying basis and reason for using cut-off scores.

##### Ratings:

Acceptable: A general discussion of the problem of setting cut-off scores and a discussion of the particular method employed is offered in the manual.

Acceptable with reservations: A very general discussion of cut-off scores is offered with little of a nature pertinent to a certain method offered.

Unacceptable: There is no discussion in the manual of the basic rationale behind the setting of cut-off scores for the test.

- F.2. Was the procedure for implementing the method explained, and was it appropriate?

##### Rationale

The test manual should contain a discussion of how the procedure for implementing the cut-off method should be used. If actual cut-off scores are given, a brief description of what they mean in terms of mastery/non-mastery decision making should be given. If only the general procedure is offered (this is not likely to be the case), then a step-by-step guide for using the procedure should be included.

Suitable methods are available for setting cut-off scores, and at least one such method should be discussed in the manual. It is proper for users to set their own cut-off scores, but suggestions for setting them should be offered in the manual.



Ratings:

Acceptable: A suitable method for establishing cut-off score(s) has been utilized.

Acceptable with reservations: A somewhat "ad-hoc" procedure has been utilized in the setting of cut-off scores. There is some discussion explaining and backing up the use of the procedure in the manual.

Unacceptable: Either an unsuitable procedure has been utilized to set cut-off scores, or no procedure at all has been utilized.

- F.3. Was evidence for the validity of the chosen cut-off score (or cut-off scores) offered?

Rationale:

If cut-off scores are offered in the manual, then some evidence for the validity of the cut-off scores should also be offered. The validity evidence should be collected during the pilot or field study, and will (most often) be assessed by relating the classification of examinees based on the particular cut-off score to some independent measure (for example, some outcome measure).

Ratings:

Acceptable: The manual presents actual data demonstrating the validity of the chosen cut-off scores, and offers a discussion about the procedure utilized.

Acceptable with reservations: The manual offers a general discussion of why the chosen cut-off scores are valid, but offers little of a concrete, substantive nature.

Unacceptable: No data or discussion is offered on this topic in the manual.

G. Validity

- G.1. Does the validity evidence offered by the test manual address adequately the intended use (or uses) of scores obtained from the test?

Rationale:

Because criterion-referenced tests are being used to determine what an examinee does or doesn't know in reference to a well-defined content area, it is crucial that the test items be content valid. Further, it should be demonstrated that the test items are construct valid for the particular use(s) for which they were intended.

Further, and more specifically, if the test scores are being used to sort examinees into mastery states, then the relationship between classifications based on test scores and some appropriately selected independent measure should be reported. (Some manuals have labeled this a reliability concern; it is not, such a relationship does not demonstrate consistency of decision-making, but rather the construct validity of the test scores.)

Ratings:

Acceptable: The manual gives a clear description of the attempts made to insure that the test is content valid. This should take the form of particular procedures used, and not be a general review of content validity. Also, a discussion of the construct validity of the scores for their intended uses should be presented, particularly when the test is used to make mastery decisions.

Acceptable with reservations: The manual contains a more general, less detailed, discussion of the content validity issue. The particulars of the procedures used to identify content and construct validity are somewhat glossed over.

Unacceptable: There is an inadequate discussion of the procedures utilized for establishing content and construct validity.

- G.2. Is an appropriate discussion of factors affecting the validity of test scores offered in the test manual?

Rationale:

With criterion-referenced tests, because examinees are not being compared with one another, there is likely to be less discussion of the factors affecting the validity of test scores, i.e., those factors that would disrupt or negate standardized testing conditions. However, if the test is being used to make individual decisions at different points in time, it is important that the testing conditions remain constant. Further, as norms tables

become more popular for use with criterion-referenced tests, standardized testing conditions are essential to insure valid comparison to the norm group. All this is to say is that the manual should contain a discussion of factor affecting the validity of test scores, and some suggestions about how to minimize such factors.

Ratings:

Acceptable: The manual contains an explicit discussion of the factors that affect the validity of test scores.

Unacceptable: The manual does not contain an explicit discussion of factors that affect the validity of test scores.

H. Norms

H.1. Are the norms data reported in an appropriate form?

Rationale:

If norms data are offered in the manual to augment the interpretability of the test scores, then the norms data should be properly reported. Any of the usual procedures for presenting norms data (percentiles, standard scores, stanines, age and grade-equivalents) can be utilized, but the procedure(s) used should fit as closely as possible to the criterion-referenced interpretations being offered.

Ratings:

Acceptable: The manual utilizes a method for presenting norms data that is useful in conjunction with the criterion-referenced interpretations. Suitable guidelines and cautions for use should also be included.

Acceptable with reservations: The manual utilizes a non-standard method of establishing norms data that is substantively correct, but is difficult to justify using. (An example is the use of regressed estimates of normative scores.)

Unacceptable: The manual explains that a norm-referenced interpretation is possible but then offers either norms tables that are difficult to use or does not offer suitable guidelines.

- H.2. Are the samples of examinees utilized in the norming study described?

Rationale:

If norms data are offered in the manual, the norms group must be clearly described. Then the potential user can see how well the norms group data fits the group of students to be tested. The user can determine whether the normative data are suitable for the use he/she has in mind.

Ratings:

Acceptable: The norms group is clearly described in the test manual when normative interpretations are offered.

Unacceptable: The norms group is not clearly described, or not described at all in the test manual.

- H.3. Are appropriate cautions introduced for proper test score interpretations?

Rationale:

The cautions offered in the manual should be two in nature, one having to do with the norms data, the other with the interpretations of the scores. In reference to the first point, cautions should be offered about what can't be done with derived scores. For instance, if percentile ranks are offered, the manual should state that such measures have ordinal properties and the units are not the same throughout the scale, meaning that percentile ranks should not be added, etc.

The second caution has to do with the test scores themselves rather than the normative or derived scores. Since criterion-referenced test scores are usually less reliable than norm-referenced scores (the tests are shorter and scores more homogeneous), the scores should be interpreted with caution when using normative data. The problem can be circumvented by using norms with grouped data, where the problem of low individual score reliability is no longer such a problem. The point made here is that the usual assumed procedure for using norms data that occur for norm-referenced tests must be approached cautiously when using a criterion-referenced test with normative data.

Ratings:

Acceptable: The manual offers suitable cautions about normative scores and the process of interpreting individual criterion-referenced test scores with normative data.

Unacceptable: Suitable cautions are not offered in the manual for interpreting scores vis-a-vis normative tables.

## I. Reporting of Test Score Information

I.1. Are the test scores reported for examinees on an objective by objective basis?

### Rationale:

The decisions usually made with a criterion-referenced test are on an objective by objective basis. Total test scores, which provide useful norm-referenced information, are not useful for determining what a student does or does not know in each content area. To determine what a student knows, data is needed on an objective by objective basis. Therefore, the test must provide such information. If the test is scores by machine, the output must be on an objective level; if the test is hand scored, suitable answer keys and record forms for each objective should be enclosed with the test.

### Ratings:

Acceptable: The test provides suitable mechanisms for reporting individual and group data on an objective by objective basis.

Unacceptable: The test does not provide suitable mechanisms for reporting data on an objective by objective basis.

I.2. Are there multiple options available to the user for reporting of test results (for example, by class and grade within a school)?

### Rationale:

Test score users often have the need for their data to be summarized in a variety of ways (class, grade, school, district, perhaps by sex). There should be sufficient options offered with the test to aid the potential user.

### Ratings:

Acceptable: Sufficient data reporting options are offered in conjunction with the test.

Unacceptable: There are not sufficient options available to satisfy the majority of the potential test users.

- I.3. Are convenient procedures available for scoring tests by hand, and forms available for reporting test score information?

Rationale:

If, for whatever reason (for example, the need to provide immediate feedback to students), the test user decides to score the test by hand, a suitable scoring key, answer form, and record form should be provided for the test. That is, the option should be made available for efficient, convenient hand-scoring of the test.

Ratings:

Acceptable: The test has included with it an answer key, answer forms, and test record form to facilitate hand scoring by the potential user.

Unacceptable: The test can't be conveniently scored by hand.

J. Test Score Interpretations

- J.1. Are suitable cautions included in the manual for interpreting individual and group objective score information?

Rationale:

The test manual should provide a discussion of the amount of error that exists in criterion-referenced test scores. In particular, a discussion of false-positive and false-negative errors that can be made when making mastery decisions is necessary. The potential user needs to know what the likelihood of his/her making a false-positive or false-negative error is if the test is used (with the given number of items per objective and given cut-off score). In a like fashion, there should be a discussion of the potential or possible error involved in using the test score as an estimate of a domain score. Finally, if group decisions are being made based on group objective scores, certain cautions should be advanced in the manual about this situation.

Ratings:

Acceptable: The manual presents an in-depth discussion of the potential problems in using criterion-referenced test scores for mastery determination and/or domain score estimation.

Unacceptable with reservations: The manual provides an overview of the problem, and offers some very general cautions.

Unacceptable: The manual presents no cautions on the use of individual and group objective scores.

J.2. Are appropriate guidelines offered for utilizing test scores to accomplish stated purposes?

Rationale:

The test manual should provide a discussion of how test scores can be used to make individual (and group) instructional decisions. Practical examples of how to go about making such decisions should be included. It is, quite simply, not enough for the test publisher to offer the test for use without appropriate guidelines for using the test scores. These guidelines will help the user in making decisions consonant with the test purpose. Without some help with the decision making, the user could end up using the test in a fashion quite different from the one for which it was intended. In particular, there should be suitable guidelines offered to aid the user in making mastery decisions. It should be clearly specified how to treat masters, non-masters, etc., in terms of decision-making, and it would be helpful if guidelines for subsequent instruction were also offered.

Ratings:

Acceptable: The manual contains suitable guidelines to aid the test user in making instructional decisions.

Acceptable with reservations: The manual gives some guidelines, but falls short of really aiding the user in the instructional decision-making process.

Unacceptable: Appropriate guidelines are not offered in the manual.

2.3.3 Development of the Evaluation Form

A shortcoming of the Standards developed by APA/AERA/NCME, and of the work of Popham (1978) and others, is the lack of a suitable evaluation form to apply a set of standards or guidelines. Such an evaluation form would be very useful, and so one was developed in conjunction with the guidelines presented in this chapter (Hambleton & Eignor, 1978a). A copy of the form is presented on the next four pages.

Criterion-Referenced Test and Test  
Manual Evaluation Form

Background Information

Test Name: \_\_\_\_\_ Forms and Levels: \_\_\_\_\_

Test Publisher: \_\_\_\_\_ Author(s): \_\_\_\_\_

Year of Publication: \_\_\_\_\_ Cost: \_\_\_\_\_

Reusable Booklets:    Yes    No

Special Test Administration Conditions: \_\_\_\_\_

Manual and Other Technical Aids: \_\_\_\_\_

Question	Ratings				Comments
	Acceptable	Unacceptable	Unsure	Not Applicable	
A.1. Is the purpose (or purposes) of the test stated in a clear and concise fashion?					
A.2. Is each objective clearly written so that it is possible to identify an "item pool"?					
A.3. Is it clear from the list of objectives what the test measures?					
A.4. Is an appropriate rationale offered for including each objective in the test?					
A.5. Can a user "tailor" the test to meet local needs by selecting objectives from a pool of available objectives?					
A.6. Is there a match between the content measured by the test and the situation where the test is to be used?					



Question	Ratings				Comments
	Acceptable	Unacceptable	Unsure	Not Applicable	
A.7. Are individuals identified who were responsible for the preparation of objectives?					
A.8. Does the set of objectives measured by the test serve as a representative set from some content domain of interest?					
B.1. Is the item review process described?					
B.2. Are the test items valid indicators of the objectives they were developed to measure?					
B.3. Is the set of test items measuring an objective representative of the "pool" of items measuring the objective?					
B.4. Are the items free of technical flaws?					
B.5. Are the test items in an appropriate format to measure the objectives they were developed to measure?					
B.6. Are the test items free of bias (for example, sex, ethnic, or racial)?					
B.7. Was a heterogeneous sample of examinees employed in piloting the test items?					
B.8. Was the item analysis data used <u>only</u> to detect "flawed" items?					
C.1. Do the test directions include information relative to test purpose, time limits, practice questions, answer sheets, and scoring?					

Question	Ratings				Comments
	Acceptable	Unacceptable	Unsure	Not Applicable	
C.2. Are the test directions clear?					
C.3. Is the test easy to score?					
C.4. Does the test manual specify an examiner's role and responsibilities?					
D.1. Is the layout of the test booklets attractive?					
D.2. Is the layout of the test booklets convenient for examinees?					
E.1. Is the type of reliability information offered in the test manual appropriate for the intended use (or uses) of the scores?					
E.2. Was the sample of examinees adequate in size, and representative of the population for whom the test is intended?					
E.3. Are test lengths suitable to produce tests with desirable levels of test score reliability?					
E.4. Is reliability information offered in the test manual for each intended use (or uses) of the test scores?					
F.1. Was a rationale offered for the selection of a method for determining cut-off scores?					
F.2. Was the procedure for implementing the method explained, and was it appropriate?					

<p>For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "✓" in the column corresponding to your answer to each question.</p> <p style="text-align: center;">Question</p>	Ratings				Comments
	Acceptable	Unacceptable	Unsure	Not Applicable	
F.3. Was evidence for the validity of the chosen cut-off score (or cut-off scores) offered?					
G.1. Does the validity evidence offered in the test manual address adequately the intended use (or uses of scores) obtained from the test?					
G.2. Is an appropriate discussion of factors affecting the validity of test scores offered in the test manual?					
H.1. Are the norms data reported in an appropriate form?					
H.2. Are the samples of examinees utilized in the norming study described?					
H.3. Are appropriate cautions introduced for proper test score interpretations?					
I.1. Are the test scores reported for examinees on an objective by objective basis?					
I.2. Are there multiple options available to the user for reporting of test results (for example, by class and grade within a school)?					
I.3. Are convenient procedures available for scoring tests by hand, and forms available for reporting test score information?					
J.1. Are suitable cautions included in the manual for interpreting individual and group objective score information?					
J.2. Are appropriate guidelines offered for utilizing test scores to accomplish stated purposes?					

#### 2.3.4 Choice of Tests for Evaluation

As a field test of the guidelines, and also as a means of identifying to the testing public the present state of standardized criterion-referenced tests, eleven of the more popular criterion-referenced tests were selected. The opinions of Dr. George Madaus, review editor for Journal of Educational Measurement, and Dr. Frank Stetz, of Harcourt Brace Jovanovich, were of great assistance in helping to choose the tests for review. It was important to be in contact with Dr. Madaus because reviews of the tests were to be published by the Journal of Educational Measurement (see Hambleton & Eignor, 1978b).

The names of the tests and some descriptive information are presented in Table 2.3.1. Each of the test publishers was contacted and it was explained to them that a review was going to be published in the Journal of Educational Measurement. Each publisher was asked to send as much relevant information as possible; the reviews were based on the information received from this request. The information for each test that was contained in the manuals, etc., was carefully read, and each test was evaluated independently of the others.

#### 2.3.5 Application of the Guidelines to the Tests

The primary purpose for evaluating the eleven tests was to ascertain the extent to which each test, and all of the tests collectively, met the guidelines. An evaluation of each test relative to each guideline was done; however, the most important information was arrived at by determining how well the tests as a group met each of the guidelines. The group information was informative because it

Table 2.3.1

## Criterion-Referenced Tests Reviewed in the Study

Code	Name of Test	Grades	Levels	Forms	Publication	
					Date	Publisher
1	1976 Stanford Diagnostic Mathematics Test	1-12	4	2	1976	Harcourt Brace Jovanovich
2	1976 Stanford Diagnostic Reading Test	1-12	4	2	1976	Harcourt Brace Jovanovich
3	Skills Monitoring System-Reading	3-5	3	1	1975	Harcourt Brace Jovanovich
4	Individual Pupil Monitoring System-Mathematics	1-6	6	2	1974	Houghton- Mifflin
5	Individual Pupil Monitoring System-Reading	1-8	8	2	1974	Houghton- Mifflin
6	Diagnostic Mathematics Inventory	1.5-7.5	7	1	1977	CTB/McGraw- Hill
7	Prescriptive Reading Inventory	K-6.5	6	1	1977	CTB/McGraw- Hill
8	Diagnosis: An Instructional Aid-Mathematics and Reading	1-6	2	2	1974	Science Research Associates
9	Mastery: An Evaluation Tool-SOBAR Reading	K-9	10	2	1975	Science Research Associates
10	Mastery: An Evaluation Tool-Mathematics	K-8	9	2	1974	Science Research Associates
11	Fountain Valley Support System in Mathematics	K-8	9	1	1974	Richard L. Zweig Associates

could be used to pinpoint areas where commercial materials were in need of revisions and further development.

In judging the quality of a test and test manual relative to each guideline, the following rating scale was used:

- A = Acceptable
- A<sup>-</sup> = Acceptable, with reservations
- X = Unacceptable, data offered was unsuitable or improperly used
- Y = Unacceptable, no data was offered
- N = Not applicable

Table 2.3.2 summarizes the ratings of the 11 tests on the 39 guidelines. What follows are some specific comments for certain of the tests. These comments were included for situations where application of the guidelines was not straightforward, and additional comments were deemed necessary.

#### Specific Comments<sup>1</sup>

##### 1976 Stanford Diagnostic Mathematics Test

- 1: The potential user can select subtests to administer, but he/she does not have the flexibility of selecting individual objectives.
- 2: Items were selected based solely on statistical properties; therefore the items are not likely to be a representative sample.

---

<sup>1</sup>The numbers for each of the comments correspond to the subscripts in Table 2.3.2.

Table 2.3.2  
Summary of Ratings of the Criterion-Referenced Tests

Question	Test										
	1	2	3	4	5	6	7	8	9	10	11
A1	A	A	A	A <sup>-</sup>	A <sup>-</sup>	A	A	A	A	A	X
A2	X	X	X	X	X	X	X	X	X	X	X
A3	A	A	A <sup>-</sup>	A	A	A	A	A	A	A	A
A4	A	A	A	A <sup>-</sup>	A <sup>-</sup>	A	A	A	A	A	X
A5	A <sup>-</sup>	A <sup>-</sup>	A	A	A	X	X	A	A	A	A
A6	A	A	A	A	A	A	A	A	A	A	A
A7	Y	Y	A <sup>-</sup>	Y	Y	Y	A <sup>-</sup>	A	A <sup>-</sup>	A <sup>-</sup>	A <sup>-</sup>
A8	A <sup>-</sup>	A <sup>-</sup>	A <sup>-</sup>	A <sup>-</sup>	A <sup>-</sup>	A <sup>-</sup>	A <sup>-</sup>	A <sup>-</sup>	A <sup>-</sup>	A <sup>-</sup>	A <sup>-</sup>
B1	X	X	A	A <sup>-</sup>	A <sup>-</sup>	X	A <sup>-</sup>	Y	A	A	Y
B2	A <sup>-</sup>	A <sup>-</sup>	A	A <sup>-</sup>	A <sup>-</sup>	? <sup>a</sup>	A <sup>-</sup>	A <sup>-</sup>	A	A	A <sup>-</sup>
B3	X <sub>2</sub>	X <sub>2</sub>	X	X <sub>1</sub>	X <sub>1</sub>	X	X	X	X	X	X
B4	A	A	A	A	A	A	A	A	A	A	A
B5	A	A	A	A	A	A	A	A	A	A	A
B6	A	A	A	Y	Y	?	Y	Y	Y	A	Y
B7	A	A	A	A	A	A	A	Y	Y	Y	Y
B8	X <sub>3</sub>	X <sub>3</sub>	A	X <sub>2</sub>	X <sub>2</sub>	X <sub>1</sub>	X <sub>2</sub>	Y	X	X	Y
C1	A	A	A	A	A	?	A	A	A <sub>2</sub>	A <sub>2</sub>	? <sup>b</sup>
C2	A	A	A	A	A	?	A	A	A	A	A
C3	A	A	A	A	A	?	A	A	A	A	A
C4	A	A	A	A	A	?	A	A	A	A	A
D1	A	A	A	A	A	?	A	A	A	A	A
D2	A	A	A	A	A	?	A	A	A	A	A
E1	A <sub>4</sub> <sup>-</sup>	X <sub>4</sub>	A <sub>2</sub> <sup>-</sup>	Y	Y	X	X <sub>3</sub>	Y	X <sub>3</sub>	X <sub>3</sub>	Y
E2	A	A	A	Y	Y	A	A	Y	A <sub>4</sub>	A <sub>4</sub>	Y
E3	A <sup>-</sup>	A <sup>-</sup>	A <sub>3</sub> <sup>-</sup>	A <sub>3</sub> <sup>-</sup>	A <sub>3</sub> <sup>-</sup>	X <sub>2</sub>	X <sub>4</sub>	X	X	X	A <sub>1</sub> <sup>-</sup>
E4	A <sup>-</sup>	A <sup>-</sup>	A <sup>-</sup>	Y	Y	X	X	Y	X	X	Y
F1	A	A	A	Y	A <sub>4</sub> <sup>-</sup>	Y	A	Y <sub>2</sub>	Y	A	Y
F2	A	A	X	Y	Y	X	X	Y	A	A	Y
F3	A	A	A <sup>-</sup>	Y	Y	Y	A <sup>-</sup>	Y	A <sup>-</sup>	A <sup>-</sup>	Y
G1	A	A	A	X	X	A	A	X	A <sup>-</sup>	A <sup>-</sup>	Y
G2	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
H1	A	A	N	N	N	A <sub>3</sub> <sup>-</sup>	A	N	N	N	N
H2	A	A	N	N	N	?	Y	N	N	N	N
H3	A	A	N	N	N	Y	Y	N	N	N	N
I1	A	A	A	A	A	?	A	A	A	A	A
I2	A	A	A	A	A	?	A	A	A	A	A
I3	A	A	A	A	A	?	A	A	A	A	A
J1	A <sup>-</sup>	A <sup>-</sup>	A	Y	Y	?	A <sub>5</sub> <sup>-</sup>	Y	A <sub>5</sub> <sup>-</sup>	A <sub>5</sub> <sup>-</sup>	Y
J2	A	A	A	Y <sub>4</sub>	Y <sub>5</sub>	?	A	A <sub>3</sub>	A <sub>6</sub> <sup>-</sup>	A <sub>6</sub> <sup>-</sup>	A

<sup>a</sup>We did not have the proper materials to assess the quality of the test in the areas marked by a "?".

<sup>b</sup>The information was on a cassette. We did not listen to the tape and so we were not in a position to rate this aspect of the test.

- 3: Norm-referenced test item analysis data alone was used to select items. The guideline calls for using criterion-referenced test item analysis data to detect "flawed" items.
- 4: An attempt was made to establish consistency of mastery-state assignment using tetrachoric correlation coefficients. While not the best approach, it is nonetheless a reasonable first approach at establishing reliability.

#### 1976 Stanford Diagnostic Reading Test

- 1: The potential user can select subtests to administer, but he/she does not have the flexibility of selecting individual objectives.
- 2: Items were selected based solely on statistical properties; therefore the items are not likely to be a representative sample.
- 3: Norm-referenced test item analysis data alone was used to select items. The guideline calls for using criterion-referenced test item analysis data to detect "flawed" items.
- 4: There is an abundance of reliability data offered for a norm-referenced usage of the scores, but none for criterion-referenced test usage.

#### Skills Monitoring System—Reading

- 1: The list of objectives is not included in the Teacher Handbook, which is included with the specimen set. The Teacher Handbook contains a list of skills statements only.
- 2: Data to be published in the future in a tech report will provide suitable reliability evidence (personal communications from publishers). For instance, consistency of mastery decisions will be studied using kappa. However, the present published reliability information is quite weak.
- 3: On the Skill Location test, there are only two items per objective and this could be a problem. The Skills-Minis, however, have eight items per objective.



Individual Pupil Monitoring System—Mathematics

- 1: It appears so, but depending upon the state of B1, the items may not be a representative sample.
- 2: Item analysis data was collected and it appears to be used to select items. The manual is unclear on this point.
- 3: It would appear so, but no data is offered. In levels 1-3, five items per objective might be a problem, but for levels 4-8, ten items per objective is sufficient.
- 4: There is a reference booklet that cross-references objectives to major texts. However, there is little in the way of providing practical classroom guidelines for using the scores from the test.

Individual Pupil Monitoring System—Reading

- 1: It appears so, but depending upon the state of B1, the items may not be a representative sample.
- 2: Item analysis data was collected and it appears to be used to select items. The manual is unclear on this point.
- 3: There are five items per objective which may be a problem, but no substantiating data is offered.
- 4: There was a rationale why Houghton-Mifflin wants individual teachers to set cut-offs. This contained some valuable information.
- 5: There is a reference booklet that cross-references objectives to major texts. However, there is little in the way of providing practical classroom guidelines for using the scores from the test.

Diagnostic Mathematics Inventory

- 1: The answer depends on how the items were selected. There is little information in the manual on that process.
- 2: There is only one item per objective. However, decisions are made using categories of objectives, usually made up of two thru eight objectives (therefore two to eight items). For certain categories, the number may still not be sufficient.

- 3: The data is discussed as being presented as regressed estimates of normative scores, but nothing more is said in the technical manual. The DMI Guide to Ancillary Materials (not provided) is probably needed.

### Prescriptive Reading Inventory

- 1: The item review process is not described in sufficient detail. The manual states only that it took place.
- 2: The item analysis data may not have been used properly. The manual says item sensitivity indices were used both to select items and to detect aberrant items.
- 3: An ad-hoc procedure involving the use of criterion tests made up of items parallel to the PRI items was used. Correlation coefficients and joint frequency distributions were utilized.
- 4: There are three or four items per objective, and it is questionable whether that is a sufficient number.
- 5: Some practical cautions are offered, but there is no discussion of false-positive and false-negative errors.

### Diagnosis: An Instructional Aid— Mathematics and Reading

- 1: It appears so, however, an overall list of objectives did not appear in the manual. Only a description of the thirty probe tests was offered.
- 2: A very weak rationale to help teachers set their own cut-offs is offered.
- 3: There is a reference booklet linking objectives to major tests. There needs to be more useful classroom guidelines to help the teachers in instruction.

### Mastery: An Evaluation Tool—Mathematics

- 1: The only item analysis data, item difficulty and item-test correlations, were collected after the test was marketed. It is difficult to ascertain the purpose of this data.
- 2: Only certain of the test levels have manuals that supply information. The manuals that exist are excellent.

- 3: The only reliability data provided is based on KR-20, which is not a suitable reliability measure when data is used to assign individuals to mastery states.
- 4: The sample was large enough, but not representative of an appropriate population.
- 5: The problem of guessing is discussed; there is no discussion of false-positive and false-negative errors.
- 6: More suggestions for classroom uses of the test scores would be helpful.

#### Mastery: An Evaluation Tool—Reading

- 1: The only item analysis data, item difficulty and item-test correlations, were collected after the test was marketed. It is difficult to ascertain the purpose of this data.
- 2: Only certain of the test levels have manuals that supply information. The manuals that exist are excellent.
- 3: The only reliability data provided is based on KR-20, which is not a suitable reliability measure when data is used to assign individuals to mastery states.
- 4: The sample was large enough, but not representative of an appropriate population.
- 5: The problem of guessing is discussed; there is no discussion of false-positive and false-negative errors.
- 6: More suggestions for classroom uses of the test scores would be helpful.

#### Fountain Valley Support Systems in Mathematics

- 1: There are three or four items per objective, probably not a sufficient number.

#### 2.4 Results and Discussion

For the potential user interested in choosing a particular test, Table 2.3.2 is most helpful in that particular strengths and weaknesses

of each test are specified. For someone interested in using the evaluations to get an impression of the present overall state of standardized criterion-referenced tests, other data would be more useful. Table 2.4.1, presents numbers and percentage of tests reviewed for each rating category on each guideline.

In reference to the guidelines involving Objectives (A.1-A.8) the following comments can be made (the relevant guideline is listed in parentheses):

1. Only about half of the publishers included information about the qualifications of individuals who prepared the objectives measured by their test. The qualifications of participants in this aspect of the test development process is important information for potential users (A.7).
2. Current commercially available "criterion-referenced tests" reviewed in this chapter should be called "objectives-referenced tests" since the tests appear to be developed from behavioral objectives (Popham, 1978). Starting to develop a test from a listing of behavioral objectives is less than ideal because behavioral objectives usually do not lead to unambiguous definitions of the "item pools" keyed to the behavioral objectives. The solution is to write "domain specifications" (Popham, 1978) (A.2).
3. Since test developers have not used "domain specifications," it is impossible to assess "item representativeness." Item representativeness is essential if users desire to use objective scores to "generalize to the domain of behavior defined by the objectives." If item representativeness is not established, scores can only be interpreted in terms of the specific items included in the test (A.8).

In reference to the guidelines involving Test Items (B.1-B.8), the following comments can be made:

1. Only three of the eleven tests described in an adequate fashion how the item review process took place. This is important information to present to potential users (B.1).
2. Only three of the eleven tests gave sufficient information that would allow the potential user to ascertain that the test items are valid indicators of the objective they were developed to measure (B.2).

Table 2.4.1

Number and Percentage of Tests Reviewed in Each Rating Category on the Guidelines for Evaluating Criterion-Referenced Tests and Test Manuals

<u>Guidelines and Categories</u>	<u>N</u>	<u>Percentage</u>
A.1 Is the purpose (or purposes) of the test stated in a clear and concise fashion?		
Acceptable	8	72.7%
Acceptable with reservations	2	18.2%
Unacceptable (X) <sup>1</sup>		
A.2 Is each objective clearly written so that it is possible to identify an "item pool"?		
Acceptable	0	0%
Unacceptable (X)	11	100%
A.3 Is it clear from the list of objectives what the test measures?		
Acceptable	10	90.9%
Acceptable with reservations	1	9.1%
Unacceptable	0	0%
A.4 Is an appropriate rationale offered for including each objective in the test?		
Acceptable	8	72.7%
Acceptable with reservations	2	18.2%
Unacceptable (X)	1	9.1%
A.5 Can a potential user "tailor" the test to meet local needs by selecting objectives from a pool of available objectives?		
Acceptable	7	63.6%
Acceptable with reservations	2	18.2%
Unacceptable (X)	2	18.2%

<sup>1</sup>X indicates that the data offered was unsuitable or improperly used.

Y indicates no data was offered.

<u>Guidelines and Categories</u>	<u>N</u>	<u>Percentage</u>
A.6 Is there a match between the content measured by the test and the situation where the test is to be used?		
Acceptable	11	100%
Unacceptable	0	0%
A.7 Are individuals identified who were responsible for the preparation of objectives?		
Acceptable	1	9.1%
Acceptable with reservations	5	45.5%
Unacceptable (Y)	5	45.5%
A.8 Does the set of objectives measured by the test serve as a representative set from some content domain of interest?		
Acceptable	0	0%
Acceptable with reservations	11	100%
Unacceptable	0	0%
B.1 Is the item review process described?		
Acceptable	3	27.3%
Acceptable with reservations	3	27.3%
Unacceptable (X)	2	18.2%
Unacceptable (Y)	2	18.2%
Unable to assess	1	9.1%
B.2 Are the test items valid indicators of the objectives they were developed to measure?		
Acceptable	3	27.3%
Acceptable with reservations	7	63.6%
Unacceptable	0	0%
Unable to assess	1	9.1%
B.3 Is the set of test items measuring an objective representative of the "pool" of items measuring the objective?		
Acceptable	0	0%
Acceptable with reservations	0	0%
Unacceptable (X)	11	100%
B.4 Are the items technically correct?		
Acceptable	11	100%
Unacceptable	0	0%

<u>Guidelines and Categories</u>	<u>N</u>	<u>Percentage</u>
B.5 Was a suitable format for the items selected?		
Acceptable	11	100%
Unacceptable	0	0%
B.6 Are the test items free of bias (for example, sex, ethnic, or racial)?		
Acceptable	4	36.4%
Unacceptable (Y)	6	54.5%
Unable to assess	1	9.1%
B.7 Was a heterogeneous sample of examinees employed in piloting the test items?		
Acceptable	7	63.6%
Unacceptable (Y)	4	36.4%
B.8 Was the item analysis data used <u>only</u> to detect "flawed" items?		
Acceptable	1	9.1%
Unacceptable (X)	8	72.7%
Unacceptable (Y)	2	18.2%
C.1 Do the test directions include information relative to test purpose, time limits, practice questions, answer sheets, and scoring?		
Acceptable	9	81.8%
Unacceptable	0	0%
Unable to assess (information should have been offered)	1	9.1%
Unable to assess (information offered, but not able to process)	1	9.1%
C.2 Are the test directions clear?		
Acceptable	10	90.9%
Unacceptable	0	0%
Unable to assess	1	9.1%
C.3 Is the test easy to score?		
Acceptable	10	90.9%
Unacceptable	0	0%
Unable to assess	1	9.1%

<u>Guidelines and Categories</u>	<u>N</u>	<u>Percentage</u>
C.4 Does the test manual specify an examiner's role and responsibilities?		
Acceptable	10	90.9%
Unacceptable	0	0%
Unable to assess	1	9.1%
D.1 Is the layout of the test booklets attractive?		
Acceptable	10	90.9%
Unacceptable	0	0%
Unable to assess	1	9.1%
D.2 Is the layout of the test booklets convenient for examinees?		
Acceptable	10	90.9%
Unacceptable	0	0%
Unable to assess	1	9.1%
E.1 Is the type of reliability information offered in the test manual appropriate for the intended use (or uses) of the scores?		
Acceptable	0	0%
Acceptable with reservations	2	18.2%
Unacceptable (X)	5	45.5%
Unacceptable (Y)	4	36.4%
E.2 Was the sample (or samples) of examinees used in the reliability study adequate in size, and representative of the population for whom the test is intended?		
Acceptable	6	54.5%
Acceptable with reservations	1	9.1%
Unacceptable (Y)	4	36.4%
E.4 Are test lengths suitable to produce tests with desirable levels of test score reliability?		
Acceptable	0	0%
Acceptable with reservations	6	54.5%
Unacceptable (X)	5	45.5%
E.4 Is reliability information offered in the test manual for each intended use (or uses) of the test scores?		
Acceptable	0	0%
Acceptable with reservations	3	27.3%
Unacceptable (X)	4	36.4%
Unacceptable (Y)	4	36.4%



<u>Guidelines and Categories</u>	<u>N</u>	<u>Percentage</u>
F.1 Was a rationale offered for the selection of a method for determining cut-off scores?		
Acceptable	5	45.5%
Acceptable with reservations	1	9.1%
Unacceptable (Y)	5	45.5%
F.2 Was the procedure for implementing the method explained, and was it appropriate?		
Acceptable	3	27.3%
Acceptable with reservations	0	0%
Unacceptable (X)	3	27.3%
Unacceptable (Y)	5	45.5%
F.3 Was evidence for the validity of the chosen cut-off score (or cut-off scores) offered?		
Acceptable	2	18.2%
Acceptable with reservations	4	36.4%
Unacceptable (Y)	5	45.5%
G.1 Does the validity evidence offered in the test manual address adequately the intended use (or uses) of scores obtained from the test?		
Acceptable	5	45.5%
Acceptable with reservations	2	18.2%
Unacceptable (X)	3	27.3%
Unacceptable (Y)	1	9.1%
G.2 Is an appropriate discussion of factors affecting the validity of test scores offered in the test manual?		
Acceptable	0	0%
Unacceptable (Y)	11	100%
H.1 Are the norms data reported in an appropriate form?		
Acceptable	3	75%
Acceptable with reservations	1	25%
Unacceptable	0	0%
Not Applicable		

<u>Guidelines and Categories</u>	<u>N</u>	<u>Percentage</u>
H.2 Are the samples of examinees utilized in the norming described?		
Acceptable	2	50%
Unacceptable (Y)	2	50%
Not Applicable	7	
H.3 Are appropriate cautions introduced for proper test score interpretations?		
Acceptable	2	50%
Unacceptable (Y)	2	50%
Not Applicable		
I.1 Are the test scores reported for examinees on an objective by objective basis?		
Acceptable	10	90.0%
Unacceptable (Y)	1	9.1%
I.2 Are multiple options available to the user for reporting of test results (for example, by class and grade within a school)?		
Acceptable	10	90.0%
Unacceptable (Y)	1	9.1%
I.3 Are convenient procedures available for scoring tests by hand, and forms available for reporting test score information?		
Acceptable	10	90.9%
Unacceptable (Y)	1	9.1%
J.1 Are suitable cautions included in the manual for interpreting individual and group objective score information?		
Acceptable	1	9.1%
Acceptable with reservations	5	45.5%
Unacceptable (Y)	5	45.5%
J.2 Are appropriate guidelines offered in the manual for utilizing test scores to accomplish stated purposes?		
Acceptable	5	45.5%
Acceptable with reservations	3	27.3%
Unacceptable (Y)	3	27.3%

---

3. In reference to whether the items measuring an objective are representative of a "pool" of items, the comments given for A.2 are again relevant here. Behavioral objectives usually do not lead to unambiguous definitions of "item pools." Without the "item pool," the user can't ascertain whether or not the test items are representative of the "pool" (B.3).
4. "Item analysis" is an area in which there are two problems: (a) too little explanation is offered for the choice of particular item statistics and of the specifics of item statistics usage, and (b) item statistics are used in test construction, thereby "biasing" the content validity of the test in unknown ways. For eight of the eleven tests reviewed item statistics were used for more than the detection of "flawed" items (B.8).

In reference to the guidelines involving Administration of the test (C.1-C.4), the following comment can be made:

1. All the tests were well constructed in this area. There were no problems with any of them.

In reference to the guidelines on Test Layout (D.1-D.2), the tests were excellent.

In reference to the guidelines on Reliability (E.1-E.4), there were a number of problems:

1. Only two of the eleven tests came at all close to providing reliability information appropriate for the intended use (or uses) of the scores (E.1).
2. There is little or no information in any of the test manuals about whether or not the test lengths are suitable to produce desirable levels of test score reliability. For six of the eleven, the test lengths appear long enough, for the other five (where there is usually four or fewer items per objective), the lengths are not sufficient. Information on how test length related to reliability would have been most helpful (E.3).
3. For tests with multiple intended uses, a few (3) did give reasonable information about reliability for one of those uses. For most, either no reliability information or inappropriate reliability information was offered (E.4).

In reference to the guidelines on Cut-Off Scores (F.1-F.3), the following comments are relevant:

1. Only five of eleven tests offered a reasonable rationale for the method offered for setting cut-offs, and for only three of the eleven was the method appropriate. Also, procedures used for setting cut-off scores are not usually explained (F.1 and F.2).
2. For only two of the eleven tests was evidence offered for the validity of the chosen cut-off scores (for example, do those examinees classified as "masters" typically perform better than "non-masters" on some appropriately chosen external criterion measure?) (F.3).

For the guidelines on Validity (G.1-G.2), the following observations can be made:

1. Most of the tests either adequately offered or attempted to offer evidence on validity of intended test score usage (G.1).
2. None of the tests gave any sort of discussion of factors affecting the validity of test scores. This is a serious shortcoming of the tests evaluated (G.2).

In reference to Norms (H.1-H.3), only four tests reported norms data, and only two of the four described the norms sample and offered cautions about the interpretations of norms data.

All the tests were assessed as acceptable on the guidelines for Reporting of Test Score Information (I.1-I.3).

Finally, in reference to Test Score Interpretations (J.1-J.2):

1. Only a few of the manuals introduced the notion of "error" in test scores. It is extremely important for users to have some indication of the "stability" of their objective scores and/or "consistency of mastery/non-mastery decisions" (J.1).
2. A number of the tests could be improved by adding sections in the manuals to aid users in utilizing test scores to make decisions (J.2).

To summarize the results reported in this section on the evaluation of the eleven standardized criterion-referenced tests selected, it seems reasonable to state that the tests are well-constructed in the non-psychometric areas (Administration, Test Layout, Reporting of Information), but most fall short of acceptability, based on the guidelines, for Reliability, Validity, and Cut-off Scores. Further, all tests, because they were developed from behavioral objectives, suffer the problem of the lack of an identifiable "item pool," thereby restricting the inferences that can be reasonably made. In defense of the test publishers, however, it should be remembered that many of the tests were published before criterion-referenced reliability and the issue of cut-off scores were suitably defined. Fortunately, an adequate technology for constructing criterion-referenced tests and using criterion-referenced test scores now exists (Popham, 1978; Hambleton et al., 1978; Hambleton & Eignor, 1978a), and hopefully the shortcomings of the tests reviewed will soon be alleviated.

## 2.5 Conclusion

This chapter had two objectives:

1. The development of a set of usable guidelines (with appropriate rationale offered for their inclusion) for use in the evaluation of criterion-referenced tests and test manuals.
2. The application of the guidelines to the evaluation of selected standardized criterion-referenced tests, and the preparation of a complete report of the results.

In reference to the first objective, 39 individual guidelines were offered for use in evaluating criterion-referenced tests and

test manuals. Included with each of the guidelines is a rationale and a rating system for their use.

In reference to the second objective, eleven of the more popular criterion-referenced tests were selected and evaluated vis-a-vis the guidelines. The tests evaluated are summarized in Table 2.3.2. In addition to evaluating each test individually, group data was presented in the form of percentage of tests reviewed on each guideline that fell into the various rating categories. Finally, a number of comments were made concerning how the eleven tests collectively measured up to each guideline and group of guidelines. It was found that the tests were well-instructed in the non-psychometric areas (Administration, Test Layout, Reporting of Information), but most feel short of acceptability for the areas of Reliability, Validity, and Cut-off Scores.

## CHAPTER III

### THE RELATIONSHIP OF TEST LENGTH TO CRITERION-REFERENCED TEST RELIABILITY AND VALIDITY

#### 3.1 Introduction

A primary concern of individuals using test scores is that the scores be both reliable and valid. While the best approaches to assessing reliability and validity are situation-specific, it is well-known that there is a direct relationship between the length of a test and the reliability and validity of the test scores. Longer tests, in general, result in test scores with better psychometric properties. For norm-referenced tests, the relationship of test length to reliability is directly expressed by the Spearman-Brown formula. In a like fashion, there is a formula that relates norm-referenced test length to the criterion-related validity of a test. However, as the discussion in the next section will demonstrate, these formulas are not appropriate with criterion-referenced tests.

For one of the two major uses of criterion-referenced tests, domain score estimation, the test length relationship to reliability and validity can be derived, and is summarized in the well-known item sampling model (Lord & Novick, 1968). It is for the other major use of criterion-referenced test scores, mastery state determination, that the necessary work is still to be done. While the literature

abounds with papers on reliability (Livingston, 1972; Swaminathan, Hambleton, & Algina, 1974; Hambleton & Novick, 1973; Huynh, 1976; Subkoviak, 1976), validity (Cronbach, 1971; Messick, 1975; Linn, 1977), and test length (Millman, 1973; Novick & Lewis, 1974; Fhaner, 1974; Wilcox, 1976) there are no published papers this author is aware of that have investigated the relationship of test length to reliability of mastery-state assignments. The only work done to date relating criterion-referenced test length to validity has been an unpublished paper by Livingston (1978), and in Livingston's paper, test length was only indirectly considered. Given the lack of research in this very important area, this author decided to pursue the topic for dissertation research. Due to the lack of empirical developments in the area, and due to the nature of the needed data, a simulation study was decided upon as the means of investigation. In this way, relevant variables to be investigated could be controlled and systematically varied, as needed.

The following three research objectives guided the work done in investigating the relationship of criterion-referenced test length to reliability and validity of mastery-state assignments:

1. Develop a computer program, relating criterion-referenced test length to several reliability and validity indices.
2. Using the program developed, conduct a simulation study relating prior distributions of ability, actual test score distributions, and loss ratios, to chosen reliability and validity indices.
3. Produce a set of tables relating test length to reliability and validity under a wide variety of simulated testing conditions.



The reliability and validity indices used in the simulation, as well as the background material upon which choices for the distributions for the simulation were made, are discussed in the next several sections.

### 3.2 Some Background Information

In this section, a number of important background considerations will be discussed. First, the reason why norm-referenced approaches to reliability and validity are inappropriate for criterion-referenced tests will be discussed. Then, the two important uses of criterion-referenced test scores, domain score estimation and assignment of individuals to mastery states will be introduced.

#### 3.2.1 Norm-Referenced Approaches to Reliability and Validity

In norm-referenced testing situations, the test user is interested in having the test spread students out along a continuum so that comparisons, such as rankings, can be done. If there is to be suitable spread of scores to facilitate ranking, then the test items comprising the test should be selected to produce a test having maximum test score variability. If all the scores tend to group, for instance, at the upper end of the score distribution (negative skew), then the needed ranking will be made quite difficult to determine. Scores will closely coincide, and because the test data contains error, any rankings made will be questionable. One individual could have the same *true score* (i.e., errorless score) as another, but be ranked higher solely because of the error.

When scores are spread, errors in the test scores have less effect on the rankings.

In criterion-referenced testing, there is little interest in making discriminations among examinees, and hence no attempt is made to select items that produce high test score variability. Interest lies in measuring, for instance, how well a student has mastered the objectives of a well-defined subject domain. If instruction is effective, as the teacher would want, then the test score variance will be small. Further, since criterion-referenced tests are usually administered before or after instruction, test score distributions tend to be homogeneous, and centered at the high or low end of the achievement scale. There will be considerable "bunching" of students at either end of the test score scale.

The difference between the purposes for norm- and criterion-referenced tests has a direct effect upon the sort of indices to be used to assess reliability and validity. Given that reliability refers to consistency of measurement, where consistency refers to making the same judgment on an individual over occasions, for norm-referenced tests a correlation coefficient serves as an excellent indication of consistency. We usually administer parallel forms of a test and correlate individuals' scores on the two occasions. If the ranking of students is unchanged over the occasions, the correlation coefficient will be +1. To the extent that there are changes in rank, the correlation coefficient will be less than one. Similarly, validity can be established by observing the relationship between test scores and an outside criterion measure. The crucial point is

that norm-referenced tests are constructed so as to insure test score variability, and this variability allows for the use of a correlation coefficient to indicate consistency of measurement, or reliability, and validity.

A correlational approach to criterion-referenced reliability and validity does not make sense. This is because such tests are not constructed or used to rank people for comparisons. Criterion-referenced tests are instead used to either ascertain how much an individual knows in reference to a content area, or domain, or to ascertain whether an individual is a master or non-master of the content area. Further, for these two uses of criterion-referenced test scores, i.e., domain score estimation and allocation of examinees to mastery states, the approaches to establishing reliability and validity differ. However, since the approaches to the former are fairly well-developed, only the latter will be considered further in this study.

### 3.2.2 Two Criterion-Referenced Test Score Uses

Regardless of which of the two uses of criterion-referenced tests you are concerned with, the following assumption always holds. We assume that the test is constructed by randomly sampling items from a well-defined, or clearly specified, domain of items measuring an instructional objective (see Popham, 1978a; Hambleton & Eignor, 1978a). If the test is to measure more than a single objective, then the items must be randomly sampled from the domain for each objective. After administering the random sample(s), there are two

basic uses that can be made of the scores. One, the score can be used as an estimate of the examinee's level of mastery on the objective. In other words, the test score can be used as an estimate of the score the student would have obtained had he/she answered all the questions in the domain. Of course, there will be error involved in using the test score as an estimate of the *domain score*, and this error can be related to the test's reliability and validity.

The other use for test scores is in assigning examinees to mastery states, where each mastery state may be keyed to a particular instructional decision. Usually, there are simply two mastery states, called mastery and non-mastery. A cut-off score is set, using any of the appropriate methods discussed in Chapter IV of this dissertation, and the individual's test score compared to the cut-off. If the score is above the cut-off, the student is assigned mastery, and moved on to study on the next objective (Hambleton, 1974). If the score is below the cut-off, the student is retained and remedial activities are usually prescribed.

One of the reasons why these two uses of criterion-referenced tests are discussed separately lies in the fact that the model and assumptions underlying each of the uses differ. The dichotomy between the two uses can best be explained by using the concept of error of measurement. We can never measure an individual's true or errorless score; there is always error in the observed test score we work with. When using the test score to estimate an examinee's domain score, error can be defined as the difference between the estimated value (the test score) and the true value (domain score).

Here the difference between test score and domain score can be conceptualized as a distance, and these differences are squared to remove the negative signs from the distances. The model relating test score to domain score can then be formulated such that the relationship minimizes these squared differences over the individuals tested. Hambleton et al. (1978), refer to such a relationship as the *squared-error loss* function. For the second use of criterion-referenced test scores, assigning or allocating examinees to mastery states, an error can occur when an examinee is assigned, based upon his/her test score, to a mastery state other than his/her true mastery state. When there are two mastery states, master and non-master, two sorts of error can occur. If the examiner estimates that the student is below the cut-off when, in fact, the student's domain score is above the cut-off, a "false-negative" error occurs. If the examiner estimates that the student is above the cut-off when, in fact, the student is not (i.e., his/her domain score is below the cut-off) then a "false-positive" error occurs (see Hambleton & Novick, 1973). Whether talking about "false-positives" or "false-negatives," the notion of error as a distance measure makes no sense.

### 3.3 Reliability and Validity for Mastery State Assignments

In this section, the reliability and validity indices used in the research reported in this chapter will be discussed.

### 3.3.1 Reliability

The diagram below will be very useful for the developments that follow. Since for this decision-oriented use of criterion-referenced test scores, reliability can be defined as consistency of decision-making across parallel forms administrations of the same test (Hambleton & Novick, 1973), a four-fold table is useful:

		Test 2		
		Master <sup>(1)</sup>	Non-Master <sup>(2)</sup>	
Test 1	Master <sup>(1)</sup>	P <sub>11</sub>	P <sub>12</sub>	P <sub>1.</sub>
	Non-Master <sup>(2)</sup>	P <sub>21</sub>	P <sub>22</sub>	P <sub>2.</sub>
		P. <sub>1</sub>	P. <sub>2</sub>	

Here's the p's refer to proportions of examinees and test 1 and test 2 can either be two test administrations of the same test, or parallel forms of a single test.

Hambleton and Novick (1973) suggested that a proportion-agreement index be used as an index of reliability. For the above situation,

$$p_o = \sum_{k=1}^2 p_{kk} = p_{11} + p_{22}$$

is the observed proportion of decisions that are in agreement. While the  $p_o$  statistic has intuitive appeal, it suffers from a limitation that the next index takes care of.

Swaminathan, Hambleton, and Algina (1974) argued that  $p_o$  does not take into account the proportion of agreement that occurs by chance alone, and therefore it could give a false impression to users of the extent of mastery classification consistency. They suggested using coefficient  $\kappa$  (Cohen, 1960) as an index of reliability. This coefficient is defined as:

$$\kappa = (p_o - p_c) / (1 - p_c)$$

where

$$p_c = \sum_{k=1}^2 p_{k.} \cdot p_{.k} = p_{1.} \cdot p_{.1} + p_{2.} \cdot p_{.2} \cdot$$

The symbols  $p_{k.}$  and  $p_{.k}$  represent the proportion of examinees assigned to mastery state  $k$  on the first and second administrations, respectively. The symbol  $p_c$  represents the proportion of agreement that would occur even if the classifications based on the two administrations were statistically independent. Thus, it can be argued that  $\kappa$  takes into account the composition of the group, and in this sense, is more group independent than the simple proportion agreement index,  $p_o$ .

In criterion-referenced testing situations, it is often the case that administering parallel forms of a test to get an estimate of  $\kappa$  is not possible. Possible reasons include: (1) extra testing would take away instructional time, and (2) only a single form of the test is available. Therefore, what is needed is a method of arriving at either  $\kappa$ , or another suitable index, based upon one administration of a test.

Subkoviak (1976) provided a procedure for estimating reliability from a single test administration; however, he preferred to work with  $p_o$  rather than  $\kappa$ . Subkoviak defined a coefficient of agreement for individual  $i$ , denoted  $p_c^{(i)}$ , as the probability of consistent mastery classification of examinee  $i$  on parallel forms, denoted  $X$  and  $Y$ . For the case of two mastery states, this probability is given by

$$p_c^{(i)} = \text{Prob}(X_i \geq c, Y_i \geq c) + \text{Prob}(X_i < c, Y_i < c), \quad [1]$$

where  $c$  is the cut-off score.  $X_i$  and  $Y_i$  are scores for examinee  $i$  on the two tests. The two terms in the equation represent the probability of examinee  $i$  being assigned to a mastery state or a non-mastery state on each test administration, respectively. The coefficient of agreement for a group of  $N$  examinees is given by

$$p_o = \frac{\sum_{i=1}^N p_c^{(i)}}{N} .$$

In order to estimate  $p_c^{(i)}$ , Subkoviak assumed that for each examinee, scores on the two forms of the criterion-referenced test were



independently and identically distributed. Further, he assumed  $X_i$  and  $Y_i$  for a fixed examinee were identically binomially distributed. This is a questionable assumption even though test item responses are usually scored 0 to 1 and item responses are independent. The assumption of a binomial model implies that the items making up the test are equally difficult and this will seldom be the case. (Fortunately, Subkoviak addressed this point in his paper and offered a substitute expression—the compound binomial model—to handle the more typical case.) With only the two assumptions above, Subkoviak was able to show

$$p(X_i \geq c) = \sum_{x_i=c}^n \binom{n}{x_i} \pi_i^{x_i} (1-\pi_i)^{n-x_i}, \quad [2]$$

and

$$p_c^{(i)} = [p(X_i \geq c)]^2 + [1-p(X_i \geq c)]^2. \quad [3]$$

Once an estimate of an examinee's domain score ( $\pi_i$ ), denoted  $\hat{\pi}_i$ , is obtained,  $p(X_i \geq c)$  can be determined by substituting  $\hat{\pi}_i$  for  $\pi_i$  in Equation [2];  $p_c^{(i)}$  is obtained by substituting the result from Equation [2] into Equation [3]. A number of possible methods could be used to estimate an examinee's domain score. Subkoviak suggested in his paper using a regressed estimate of  $\pi_i$ , but the merits of this approach would depend on the sample estimates of group mean performance and reliability (as he correctly noted). He also offered a number of other possible domain score estimates, several of which have been reported by Lord and Novick (1968). Finally, a group estimate of the expected proportion of agreement can be

obtained by averaging the values of  $p_c^{(1)}$ , for  $i=1, 2, \dots, N$ , where  $N$  is the number of examinees in the group.

Subkoviak's approach to estimating the consistency of mastery classifications across parallel-form administrations can provide either individual or group information, and can be estimated from a single administration of a test. The only two minor problems are that the probability estimates are inflated due to the inclusion of chance agreement and that it is unreasonable to assume all items in a criterion-referenced test are equally difficult. However, on this latter point, Subkoviak has also offered a slightly different model (compound binomial) which is capable of handling the situation.

Subkoviak's method makes it possible to compute the coefficient of agreement in mastery status across occasions for an individual, and also the coefficient of agreement for a group of  $N$  persons. Since the formulas developed by Subkoviak are somewhat complex, a step-by-step procedure will be specified.

The steps in the method are as follows:

1. Obtain an estimate of the proportion of items in the whole domain of items an examinee can answer correctly. A convenient estimate is obtained by setting  $\hat{\pi}_i = \frac{x_i}{n}$ , where  $\hat{\pi}_i$  = proportion-correct score for examinee  $i$ ,  
 $x_i$  = his/her test score,  
 $n$  = total number of items included in the test (measuring the objective of interest).
2. Determine the probability that the examinee's score is greater than or equal to the cutting score ( $c$ ) using the form of the underlying (*binomial*) distribution. The probability is given by:

$$P(x_i \geq c) = \sum_{x_i=c}^n \binom{n}{x_i} \hat{\pi}_i^{x_i} (1-\hat{\pi}_i)^{n-x_i}$$

where  $\hat{\pi}_i$ ,  $x_i$ ,  $c$  and  $n$  are defined as before, and

$$\binom{n}{x_i} = \frac{n!}{x_i!(n-x_i)!}$$

where  $n! = n(n-1)(n-2) \dots$

- Using the result from step (2), compute the *coefficient of agreement for person i* using the following formula:

$$p_c^{(i)} = [P(x_i \geq c)]^2 + [1 - P(x_i \geq c)]^2$$

- Finally, compute the coefficient of agreement  $p_c$  for a group of  $N$  persons, using the following formula:

$$p_c = \frac{\sum_{i=1}^N p_c^{(i)}}{N},$$

which is the mean of the individual coefficients.

The final result,  $p_c$ , provides an estimate of the *coefficient of agreement* for the group had two test administrations taken place. The subscript  $c$  is included to clarify that the coefficient is dependent upon the assigned cutting score. If  $p_c$  is high (re: close to one), we can be sure that there would be a high degree of consistency of placement into mastery states over the two occasions.

If the number of test items is small, sometimes a better estimate (than  $\hat{\pi}_i$ ) of an examinee's domain score can be obtained by using a regressed estimate of domain score [ $\hat{\pi}_i = \hat{\pi}_i r + \bar{\pi}_i (1-r)$ , where  $r$  = test reliability, and  $\bar{\pi}_i$  = average proportion-correct score for the examinees]. The "improved" estimate can be substituted for  $\hat{\pi}_i$  in step 2. A convenient way to estimate  $r$ , the test reliability,

is to use Kuder-Richardson formula—21 ( $KR_{21}$ ). The estimate  $\hat{\pi}_1$  is then used in step 2 and the remainder of the steps follow as before.

In summary, three reliability indices were used to study the test length-reliability relationship: (1) proportion agreement, (2) coefficient kappa, and (3) Subkoviak's one administration estimate of proportion agreement.

### 3.3.2 Validity

When using a criterion-referenced test to make mastery/non-mastery decisions, the concept of validity is relatively easy to formulate. One wants the mastery/non-mastery decisions made on the basis of the test results to coincide with the decisions made had all the items in the domain been administered. For an individual, if all the items in a domain were administered to him/her, a "true decision" about mastery would be made. This is not possible; hence, a sample from the domain (the test) is administered. The test scores are said to be valid to the extent that the decision made with test scores coincide with the true decision. Further, as with reliability, there is no formula that relates criterion-referenced test length to an index of validity. Different combinations of relevant variables (test length, observed cut-off, true-mastery level cut-off) need to be observed in reference to their effects on validity indices.

When one works with real data, very seldom does the whole domain of items exist so that "true-decisions" may be determined.

A possible exception where the whole domain may exist is with the use of *item-forms analysis* (Hively et al., 1973). For most situations, however, the whole domain of items cannot be specified. This greatly increases the utility of a computer simulation approach to the problem. With simulated data, true or domain score is known, and in turn, the domain score can be referred to the true mastery cut-off, and true mastery status determined. Test data can then be simulated, referred to the observed mastery cut-off, and mastery status determined. The results can then be compared, using any of a number of indices presented in the next paragraph. Using a two-fold contingency table, the situation can be depicted as follows:

		True Status		
		Master (1)	Non-Master (2)	
Test Score (Observed)	Master(1)	$P_{11}$	$P_{12}$	$P_{1.}$
	Non-Master(2)	$P_{21}$	$P_{22}$	$P_{2.}$
		$P_{.1}$	$P_{.2}$	

(p's are again proportions)

The situation is very similar to that for reliability, and hence certain of the approaches and related indices used are similar. In fact, of the six to be discussed, one of them, the proportion agreement index is exactly the same.

A proportion-agreement index,  $p_o$ , can be calculated as follows:

$$p_o = \sum_{k=1}^2 p_{kk} = p_{11} + p_{22} \ .$$

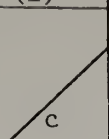
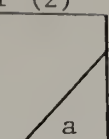
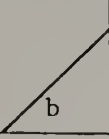
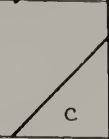
In a like fashion, although it is not typically done, a proportion of disagreement index can be calculated:

$$p = p_{21} + p_{12} = 1 - p_o \ .$$

This will be referred to later in this discussion.

Berk (1976) has suggested using the validity coefficient for studying cut-off scores and reporting test score validity information. The development that follows is from Berk (1976); necessary adaptations have been made to fit this context.

The validity coefficient being discussed is actually the Pearson correlation coefficient between the two dichotomous variables established from mastery/non-mastery classification on the test sample and on the domain being considered; each student is assigned a 1 for mastery on the test, a 0 for non-mastery; likewise for the domain. A high correlation then indicates a high probability of correct decisions on the test; or a low probability of "false-positives" and "false-negatives." Recalling, the contingency table presented earlier, the validity or phi coefficient  $\phi$  can be easily computed.

		True Status		
		Master (1)	Non-Master (2)	
Test	Master (1)	$P_{11}$ 	$P_{12}$ 	$P_{1.}$
	Non-Master (2)	$P_{21}$ 	$P_{22}$ 	
		$P_{.1}$	$P_{.2}$	

$$\phi = \frac{P_{11} - (P_{.1})(P_{1.})}{\sqrt{(P_{.1})(1-P_{.1})(P_{1.})(1-P_{1.})}}$$

Only when  $p_{1.} = p_{.1}$  can the maximum value of the coefficient be reached ( $\phi = 1.0$ ).

When looking at the proportion disagreement index presented on the previous page, false-negative and false-positive errors are weighted equally in the index. In most objectives-based programs that utilize criterion-referenced tests, equal weighting is questionable. It is usually much more serious to pass a student whose true score is below the cut-off (false-positive error) than it is to retain a student who is in reality a master (false-negative error). While the latter student may suffer boredom, the former will suffer a loss of instructional time and also perhaps experience motivational problems. Also, when the proportion-agreement index,  $p_o$ , is calculated, correct identification of masters and non-masters are weighted equally. In certain situations, this may not likely be the optimal weighting procedure. The gain in efficiency may be greater for the correct identification of masters than for correct identification of non-masters. Obviously, what is needed is a

procedure that allows for the unequal weighting of the four cells in the table above. Of course, the decision on the weights to use is judgmental. Berk (1976) suggests the weights be set based on an examination of factors involved in the decision. According to Berk: "Factors which may affect classification decisions are student motivation, teacher time, availability of instructional materials, cost of materials, duration of the instructional sequence, content of the instruction, and the specific objectives to be mastered" (p. 7).

There are two ways of proceeding in weighting the cells; both the procedures fall under the rubric of *utility analysis*. In one procedure, all four cells are weighted. (See the previous figure for the assigned weights, denoted a, b, c, and d.) The weights a and b for the incorrect decisions are negative and can be equal or unequal. Equal or unequal weights c and d are positive. Following the formulation of Berk (1976), expected utility ( $v$ ) and expected disutility ( $\delta$ ) can be calculated.

$$v = c p_{11} + d p_{22}$$

$$\delta = b p_{21} + a p_{12} \quad .$$

An overall index of utility ( $\gamma$ ) is given by

$$\gamma = v + \delta \quad .$$

The second procedure assumes that only the false decisions are weighted, and the weights are usually positive. In this case, an index of utility ( $u$ ) is given by:

$$u = b p_{21} + a p_{12} \quad ,$$



and the smaller the index the better. Thus, if a number of two-fold decision tables were being compared, and the weights  $a$  and  $b$  were the same across tables, the procedure with the maximum utility, and the one to choose, would have the smallest  $u$  value.

Livingston (1978) has developed two indices of efficiency based on linear utility. According to Livingston (1975):

Utility functions of this form imply that the cost of a bad decision is proportional to the size of the error. Similarly, they imply that the benefit from a good decision is proportional to the size of the error that was avoided. (p. 4)

Utility functions then don't assume that all false decisions are equally as serious; a threshold loss function approach does. Rather, according to Livingston (1978):

Linear utility implies that if Jones' true score is ten points above the pass/fail cut-off, while Smith's true score is five points above the cut-off, then failing Jones is twice as serious an error as failing Smith. (p. 5)

The first index weights false-positive and false-negative errors equally, while the second allows for unequal weighting. The two formulas are:

$$EFF = \frac{\sum_{t=1}^N (t-t^*) \text{ sign } (x-x^*)}{\sum_{t=1}^N (t-t^*) \text{ sign } (t-t^*)}$$

$$WEIGHTEFF = \frac{\sum_{\substack{N\exists: \\ t < t^*}} (t-t^*) \text{ sign } (x-x^*) + k \sum_{\substack{N\exists: \\ t > t^*}} (t-t^*) \text{ sign } (x-x^*)}{\sum_{\substack{N\exists: \\ t < t^*}} (t-t^*) \text{ sign } (t-t^*) + k \sum_{\substack{N\exists: \\ t > t^*}} (t-t^*) \text{ sign } (x-x^*)}$$

where  $t$  = true score,  $x$  = observed score,  $t^*$  and  $x^*$  are the true and observed cut-off points. Further.

$$\text{sign } (x-x^*) = \begin{cases} +1 & \text{if } (x-x^*) > 0 \\ -1 & \text{if } (x-x^*) < 0 \end{cases}$$

$$\text{sign } (t-t^*) = \begin{cases} +1 & \text{if } (t-t^*) > 0 \\ -1 & \text{if } (t-t^*) < 0 \end{cases}$$

$$\text{and } k = \frac{a}{b}$$

where  $a$  and  $b$  are defined as they were in the explanation of the utility coefficients.

A look at unweighted efficiency, by far the simpler of the two formulas, is clarifying. The numerator and denominator sum across true or domain scores for the  $N$  individuals. If for each individual the decisions on observed and true score coincide, i.e.,  $\text{sign } (x-x^*) = \text{sign } (t-t^*)$ , then maximum efficiency = 1.00 is obtained. To the extent that for certain individuals, these terms don't agree, the index will be less than one.

In summary, six indices are used to examine the validity question. These include: (1) proportion agreement, (2) a validity or phi coefficient, (3) four-fold utility, (4) two-fold disutility, (5) Livingston's unweighted efficiency, and (6) Livingston's weighted efficiency.

### 3.4 Criterion-Referenced Test Length

The research done to date on criterion-referenced test length (Millman, 1973; Novick & Lewis, 1974; Phaner, 1974; Wilcox, 1976) has looked at the relationship of test length to misclassification errors. Misclassification errors are of two types: false-positive errors, which occur when non-masters are assessed as masters based on test results; and false-negative errors, which occur when masters are assessed non-mastery status on a test. The longer the test is, the less the chance there is of making classification errors. However, practicality dictates against having long tests, due to time problems, construction problems, etc. Thus, the problem becomes one of determining what minimal test length is sufficient to ensure that classification errors will not exceed some specified level.

Millman (1973) considered the error properties of mastery classification decisions made by comparing a domain score estimate to an advancement score. By introducing the binomial test model, it is simple to determine the probability of misclassification, conditional upon an examinee's domain score, an advancement score, a cut-off score, and the number of items in the test. (An advancement score is distinguished from a cut-off score in Millman's work in the following way: The advancement score is the minimum number of items that an examinee must answer correctly to be assigned to a mastery state. The cut-off score is the point on the domain score scale used to separate examinees into true mastery and true non-mastery states.) By varying test length and the advancement score,

an investigator can determine the test length and advancement score that produces a desired probability of misclassification for a *given* domain score.

By making the following assumptions, Millman was able to obtain a solution to the test length problem:

1. The test is a *random* sample of dichotomously scored (0-1) items from the domain,
2. The likelihood of correct response is a fixed quantity across all test items for an individual.
3. Responses to questions on the test are independent, and
4. Errors fit the binomial test model.

No assumptions involving item content or difficulty are necessary, nor are any group based indices used.

The primary problem in applying Millman's procedure (1973) is that one would need to have a good prior estimate of an examinee's domain score. Other problems have been suggested by Novick and Lewis (1974). They reported that for certain combinations of cut-off scores and test length, changing one or both to decrease the probability of misclassification for those above the cut-off score will actually increase the probability of misclassification for those below the cut-off score. In order to choose the appropriate combinations of test length and advancement score, one must have some idea of whether the preponderance of examinees are above or below the cut-off score and one must have some idea of the relative costs of misclassification. However, the first requirement can only be satisfied with prior information about the domain scores of the group of examinees.

Novick and Lewis (1974) suggested that it would be useful to have some systematic way of incorporating prior knowledge into the test length determination problem. Further, instead of considering the probability that a student will attain a test score, given his/her true level (an unknown), it would be better to consider the probability that a student's true score exceeds a given cutting score, given his/her test score. A student will then be passed on to the next unit only if there is a sufficiently high probability that his/her true score exceeds the cutting score, given his/her test score. The procedures offered by Novick and Lewis allow such a probability to be assessed.

Millman (1973) and Novick and Lewis (1974) have prepared extensive sets of tables to use with their procedures for determining test length. Novick and Lewis' tables are particularly useful in that they allow the user to see the effects of different prior distributions, different weightings of the misclassification errors, and different expected values of the prior distributions, on test lengths and cut-off scores.

Fhaner (1974) and Wilcox (1976) also relate test length to misclassification errors, but their underlying approach is somewhat different from Novick and Lewis' (1974) and Millman's (1973). The authors do use the binomial distribution, but they look at errors of misclassification through the use of an indifference zone. The discussion that follows merges the work of Fhaner and Wilcox, using Wilcox's notation. In what follows, the binomial

distribution is used to estimate the probability of an examinee whose domain score is  $\pi$  obtaining a test score of  $x$  items out of  $n$  items.

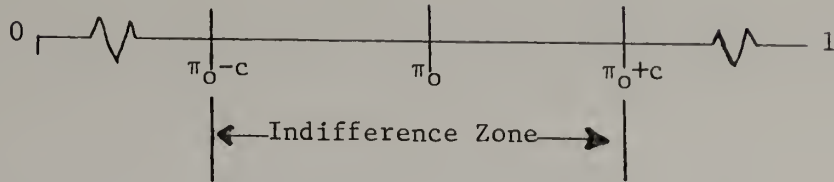
Once again, tests are used in a context; the context for criterion-referenced testing is decision making, where the test score will be used to classify individuals. To separate individuals into mastery states, a cutting score  $\pi_0$  is established such that if  $\pi < \pi_0$  the examinee is a non-master; if  $\pi \geq \pi_0$  the examinee is a master. The tester has only the test score  $x$  to work with, not  $\pi$ , and needs to decide if  $\pi < \pi_0$  or  $\pi \geq \pi_0$ . Hence, there is the risk of false-positive errors ( $\pi < \pi_0$ , but the examinee passes the test) or false-negative errors ( $\pi \geq \pi_0$ , but the examinee fails the test). Let  $\alpha$  be the probability of a false-positive error and  $\beta$  be the probability of a false-negative error. A cut-off score,  $n_0$ , needs to be established such that:

$$\begin{aligned} \text{Prob}(x \geq n_0 | \pi) &< \alpha \text{ for all } \pi < \pi_0 \\ \text{Prob}(x < n_0 | \pi) &< \beta \text{ for all } \pi \geq \pi_0 . \end{aligned}$$

Since  $\alpha = 1 - \beta$ , it is not possible to keep both probabilities at acceptably low levels. An explicit solution to the problem is generated by establishing an indifference zone. Let  $c$  be a positive constant, and form the open interval  $(\pi_0 - c, \pi_0 + c)$ . For individuals whose true score is close to  $\pi_0$  (within the interval from  $\pi_0 - c$  to  $\pi_0 + c$ ), we are "indifferent" as to how they are classified, re: there is negligible loss in misclassification of such individuals.

For individuals whose true scores are greater than  $\pi_0 + c$  or less than  $\pi_0 - c$ , we want to be reasonably certain correct decisions are made. Schematically,

Domain Score [0, 1]



Thus far we have been working with the domain of tasks. We must now specify procedures involving the test itself. Let  $n_0 =$  the cut-off or advancement score on the test. Thus, if  $x \geq n_0$ , the student is advanced; if  $x < n_0$ , the student is retained. A correct decision is made for the student if  $x < n_0$  and  $\pi < \pi_0$  or  $x \geq n_0$  and  $\pi \geq \pi_0$ . Let  $P^*$  be a number such that  $\frac{1}{2} < P^* < 1$ . Our goal is to establish  $n$  as small as possible (for a certain  $n_0$ ) so that for values of  $\pi$  not in the indifference zone, the probability of a correct decision is at least  $P^*$ .

For values of  $\pi < \pi_0 - c$ , the minimum probability of a correct decision occurs at the point  $\pi_0 - c$  and is given by

$$\alpha = \sum_{x=0}^{n_0-1} \binom{n}{x} (\pi_0 - c)^x (1 - \pi_0 + c)^{n-x} .$$

For values of  $\pi \geq \pi_0 + c$ , the minimum probability of a correct decision occurs at the point  $\pi_0 + c$  and is given by:

$$\beta = \sum_{x=n_0}^n \binom{n}{x} (\pi_0 + c)^x (1 - \pi_0 - c)^{n-x} .$$

Now to choose  $n$ , Wilcox specifies:

In particular, we choose the smallest integer  $n$  so that  $\alpha$  and  $\beta$  are greater than or equal to  $P^*$  which implies the probability of a correct decision is at least  $P^*$  for  $\pi \geq \pi_0 + c$  and  $\pi < \pi_0 - c$ . (p. 361)

Wilcox provides tables for various combinations of the variables involved in the formula. In order to use these tables, the following must be specified:

1.  $\pi_0$ : The cutting score for the domain of items. (Wilcox specifies the  $\pi_0$ 's to be .70, .75, .80, .85.)
2.  $c$ : The positive constant that forms the indifference zone. (Wilcox uses  $c = .05$  and  $c = .10$ . Thus, for  $\pi_0 = .75$  and  $c = .10$ , we are indifferent as to our classification for scores in the interval [.65, .85].)
3.  $P^*$ : The minimum probability of a correct decision for scores not in the indifference region. (Wilcox uses  $P^* = .75$ .)

By specifying these values, Wilcox's table then gives you  $n$  and  $n_0$ , along with the probability of correctly classifying examinees with true scores  $\geq \pi_0 + c$  or  $< \pi_0 - c$ .

Merging the work of Fhaner and Wilcox, the following comments can be made:

1. If  $c = 0$ , that is, there is no indifference region, it is not always possible to choose  $n$  such that the probability of a correct decision is at least  $P^*$ . Wilcox says that for this situation the probability of a correct decision approaches .5 (an unacceptable level) as  $n$  increases. Hence, Millman's solution may not be adequate for certain situations.
2. If the loss in misclassifying an individual who has obtained mastery ( $\pi \geq \pi_0 + c$ ) is different from the loss in misclassifying a non-master ( $\pi < \pi_0 - c$ ), then two numbers  $P_1^*$  and  $P_2^*$  can be chosen such that  $\frac{1}{2} < P_1^* < 1$  and  $\frac{1}{2} < P_2^* < 1$  and there is a *smallest* integer  $n$  so that  $\alpha \geq P_1^*$  and  $\beta \geq P_2^*$ .



3. If  $n$  is large, the Central Limit Theorem justifies the use of the *normal distribution* in place of the binomial. In this case, tables of the normal distribution function may be used, and use of the Wilcoxon tables can be circumvented. In this case, the number of test questions is given by:

$$n = \left( \frac{Z_{1-\alpha} \sqrt{(\pi_0 - c)(1 - \pi_0 + c)} + Z_{1-\beta} \sqrt{(\pi_0 + c)(1 - \pi_0 - c)}}{2c} \right)^2$$

where  $n$  = number of items

$c$  = positive constant (same as before)

$\pi_0$  = cutting score for domain

$Z_{1-\alpha}$  = deviation score in a *standardized normal* distribution corresponding to  $1-\alpha$

$Z_{1-\beta}$  = deviation score in a *standardized normal* distribution corresponding to  $1-\beta$ .

Fhaner notes that the normal approximation underestimates the number of items needed. Wilcoxon notes that the procedure does not give you an optimal  $n_0$  (i.e., advancement score). Hence, a user needs to be careful when making use of the normal approximation.

#### 3.4.1 Hsu's Study of Test Length-Reliability

Hsu (1977) has tied together in an unpublished paper the work of Subkoviak (1976) on reliability and Wilcoxon (1976) on test length to formulate a procedure for determining test length to satisfy minimum reliability standards. Hsu begins by formulating Subkoviak's proportion agreement index for person  $i$  in the proportion metric.

That is:

$$p_c^{(i)} = [\text{Prob}(\hat{\pi}_i \geq \pi_c)]^2 + [1 - \text{Prob}(\pi_i \geq \pi_c)]^2$$

where  $\hat{\pi}_i$ , is an estimate of  $\pi_i$ , such as  $\hat{\pi}_i = \frac{x_i}{n}$ , and  $\pi_c$  is the cut-off score. Then, using Wilcoxon's indifference zone approach,

there is a zone of small width surrounding  $\pi_c$ , whose endpoints are  $\pi_L$  and  $\pi_U$ , that separates the continuum into three areas. Within the zone, errors of misclassification are not considered serious. That is, the tester is indifferent if a true non-master is classified as a master or a true master is classified as a non-master. Outside the range, the tester is concerned about and wants to minimize these errors.

Given  $\pi_L$  and  $\pi_U$ , Hsu points out that  $\pi_c$  can be located as the value that equates the risk of misclassifying a person whose true score is  $\pi_U$  with that of a person whose true score is  $\pi_L$ . In this case,  $\pi_c$  is not likely to fall midway between  $\pi_L$  and  $\pi_U$ , and the point can't be exactly located. Hsu applies an arc sin normalizing transformation to the data, given by  $\theta_i = 2 \arcsin \sqrt{\pi_i}$ . In this metric,  $\theta_c$  falls midway between  $\theta_L$  and  $\theta_U$ . Using this procedure, Hsu is concerned about the reliability of the criterion-referenced test for those people whose true score is outside the zone  $(\pi_L, \pi_U)$ , or using the arc sin transformation,  $(\theta_L, \theta_U)$ . The *minimum* possible reliability for these persons occurs when their  $\pi_i$ 's (or  $\theta_i$ 's) equal  $\pi_L$  or  $\pi_U$  ( $\theta_L$  or  $\theta_U$ ). By applying the indifference zone procedure, Hsu is able to write Subkoviak's formula as

$$p_c^{(1)} = [\text{Prob}(\theta_i \geq \theta_c)]^2 + [1 - \text{Prob}(\theta_i \geq \theta_c)]^2$$

in the new  $\theta$  metric. However, the  $\theta_i$ 's are normally distributed and this simplifies the formula greatly. Letting  $D = \theta_U - \theta_L$ , and  $Z$  designate the standardized normal deviate, the minimum reliability

of the test (for individual  $i$ , located at either  $\theta_L$  or  $\theta_U$ ) is given by:

$$p_{\min}^{(i)} = [\text{Prob } (Z > \frac{1}{2} D \sqrt{n})]^2 + [1 - \text{Prob } (Z > \frac{1}{2} D \sqrt{n})]^2$$

Hsu then states that for all people outside the indifference zone, the minimum reliability is given by the previous formula. Since the coefficient of agreement for the group is the mean of the coefficients of each individual in the group, and each individual has the same  $p_{\min}^{(i)}$ , then this coefficient must equal:

$$p_{\min} = [\text{Prob } (Z > \frac{1}{2} D \sqrt{n})]^2 + [1 - \text{Prob } (Z > \frac{1}{2} D \sqrt{n})]^2$$

Thus, for this research done by Hsu, minimum reliability as estimated by Subkoviak's method, depends only on  $D$ , the size of the indifference zone ( $D = \theta_U - \theta_L$ ) and  $n$ , the number of items in the test. For a fixed size indifference zone, the formula gives a direct relationship between test length and minimum possible reliability.

While Hsu's research is an excellent start in the direction of formulating ways of relating test length to criterion-referenced test reliability, there are a number of shortcomings concerning the research and the underlying assumptions of Hsu's work. Of the five comments that follow, the first three are directly related to the research reported in this chapter. These three comments are:

1. Subkoviak's and Wilcox's procedures, which Hsu utilizes, are based on the binomial model. This implies that the probability of a correct response remains constant across items, or that the items are equally difficult, if the model holds. This is not likely to be the case; both authors are aware of this and suggest use of the compound binomial model instead (Lord, 1965). Wilcox (1977, 1978)

has recently utilized the compound binomial model in his work on estimating the likelihood of false-positive and false-negative errors and in estimating true score, but it has yet to be related to reliability.

2. Hsu's approach assumes equal losses for misclassifying an examinee who has attained mastery (true master classified as non-master on the test) and one who has not (true non-master classified as master in the test). Because there are numerous instances where the losses are unequal, procedures need to be developed for the test length-reliability issue that reflect this fact.
3. The research cited gives *highly conservative estimates*. For instance, for a test to discriminate between persons who can answer more than 80% of the items on the domain of items from those who can answer less than 67%, and have a minimum reliability, of .80, 64 items are needed. The conservative results are obtained because in the formula derivation, all persons are assumed to be at either  $\pi_L$  or  $\pi_U$ , the two places on the ability continuum where misclassification errors will be a maximum. Practitioners are highly unlikely to construct tests with 64 items to measure a *single* objective. Needed are tables that do not provide such conservative results. This could be done if practitioners were trained to specify a prior distribution of test score performance, and formulations were available to handle the new information.

The other two comments that can be made concerning Hsu's research are relevant, but had little or no bearing on the research reported here. These are:

1. The research cited above depends upon the use of the following arc sin normalizing transformation

$$\hat{\theta}_i = 2 \text{ arc sin } \sqrt{\alpha_{21} \left(\frac{x_i}{n}\right) + (1 - \alpha_{21}) \left(\frac{M_i}{n}\right)}, \text{ where}$$

the quantity inside the radical is the Kelly regressed estimate of  $\hat{\pi}_i$ . This differs from the frequently used Freeman-Tukey normalizing transformation

$$\hat{\theta}_i = \frac{1}{2} \left( \sin^{-1} \sqrt{\frac{x_i}{n+1}} + \sin^{-1} \sqrt{\frac{x_i+1}{n+1}} \right)$$

(Novick, Lewis, and Jackson, 1973). It needs to be determined which of the normalizing transformations is optimal for the usual criterion-referenced data conditions, that is, high  $\hat{\pi}$  values and small  $n$ .

2. Hsu's research relates the setting of an indifference zone to the reliability formula due to Subkoviak (1976). Huynh's (1976) single administration approach to reliability, based on the beta-binomial model, also warrants investigation. While harder to work with computationally, Huynh's procedure is perhaps more theoretically justifiable.

### 3.5 Research Methodology

In section 3.3, the indices of reliability and validity used in this study were presented. In this section, the other variables under control of this researcher, along with the procedure used for generating tables relating test length to the relevant indices of reliability and validity, will be presented. However, before discussing the methodology further, some comments need to be made about the use of simulation procedures, both in research settings in general, and more importantly, in terms of the research presented here.

Simulation procedures, as a mode of research investigation, have both positive and negative features. While researchers in a field will agree that well-planned empirical studies to investigate the variable or variables in question is the preferred method of investigation, this is not always possible. Any time a number of other variables must be controlled to observe the variable(s) of interest, empirical procedures become questionable. No matter how well planned the empirical research, if the other variables can't

be controlled properly, interpretative problems occur. Simulation procedures offer a practical way out of this problem in that the other variables that effect the variable(s) under study can be controlled. Further, simulation procedures are reasonable when research in a field is in the initial stages. The relationships among variables can be observed and then used as a basis for the direction and focus of later empirical studies.

Research relating test length to reliability and validity indices is in an "infant" stage. As discussed in the previous section, research has been done on the test length issue alone and on reliability and validity, but very little has been done merging the two areas. Further, due to the nature of the research, a number of different variables, such as cut-off score, prior true mastery distribution, test length, and weighting of classification errors, need to be considered in terms of their effect on reliability and validity indices. Finally, the investigation of validity requires that an examinee's true mastery status be known, that is, whether or not the individual is above or below a true mastery cut-off. Because of these three considerations: (1) infancy of the area of research, (2) the number of variables to be manipulated, and (3) the need to know true mastery status, it was decided that a simulation study was, at present, the only reasonable way to approach this research. Hopefully, the results of this simulation study will provide a basis for the design of some empirical studies to further investigate some of the more interesting results presented.

In this simulation, individuals are "sampled" from mastery distributions specified in advance, and their performance simulated on two tests. The number of items on the two tests is equal, but this number was varied from 2 to 40 items. Simulated examinee performance on the two tests was then used to investigate reliability for the various test lengths. Further, because the true mastery level of the sampled examinees was known, the relationship of observed performance to true mastery can be observed for tests of different length.

### 3.5.1 Variables Under Investigation

#### Test Model

Both the binomial and compound binomial models were used to generate simulated test performance data. While criterion-referenced test data has often been assumed to fit the binomial model, Lord (1965), and more recently, Wilcox (1976, 1977), have suggested that the compound binomial model may be appropriate. The binomial model assumes that the probability of a correct response for an individual is the same across all items on a test; or alternatively, that all items are equally difficult (for that individual). The compound binomial model assumes that the probability of correct response for an individual varies across items in a test, or that the items are not equally difficult. Investigations that have utilized both models (for instance, Subkoviak, 1976) have demonstrated some, but not drastically different results from the use of the two models. Both models were used in the simulations.

### Prior Distributions

For the binomial model, either a user-supplied or a beta prior distribution on domain scores is specified and individuals sampled from this distribution. For the user-supplied prior, a percentage of respondents is assigned to each of ten equal intervals from 0.00 to 1.00, and a distribution constructed from this information. The percentages assigned to the intervals reflect the user's belief about the domain score distribution for a relevant group of examinees. An individual is then sampled from this prior distribution, and his/her associated domain score used to simulate binomial model test performance. This process is then repeated across individuals.

When the prior distribution on mastery is specified as a beta prior, the fractile assessment procedure (Novick & Jackson, 1974) is used to specify the parameters of the beta distribution, and then a IMSL Subroutine (GGBTA) used to generate the distribution. The justification for using a beta distribution stems from two facts. One, the beta distribution is defined on a 0 to 1 interval, whereas most of the other distributions considered, such as the normal, are not. The 0-1 interval is important in that it can be directly linked to a domain score for an individual. Second, the beta distribution allows the user to easily generate skewed distributions of domain scores to approximate distributions that might be expected to occur with real test data.

The fractile assessment procedure (FASP) has been offered by Novick and Jackson (1974) as a means for specifying the parameters of a beta distribution. The user is asked to specify  $q_1$ ,  $q_2$ , and



$q_3$ , the first, second (median), and third quartiles of the distribution. The parameters,  $a$  and  $b$ , of the beta distribution are then (approximately) given by:

$$a = cq_2 + \frac{1}{3} \text{ and } b = c(1 - q_2) + \frac{1}{3}$$

where

$$c = .057 \left( \frac{1}{d_1} + \frac{1}{d_3} \right)$$

where

$$d = \left[ [q_2 (1-q_3)]^{\frac{1}{2}} - [q_3 (1-q_2)]^{\frac{1}{2}} \right]^2 .$$

The parameters  $a$  and  $b$  are then used as input to the GGBTA Subroutine, which generates (internally) a beta distribution and samples an individual's domain score from the distribution. As with the "user-supplied" prior distribution, this domain score is then used to simulate binomial model test performance. This process is then repeated across individuals.

When data were simulated to fit the compound binomial model, a different procedure was used. Rather than specifying a prior domain score distribution, and then sampling from that distribution, a more complex procedure was necessary. This is because for the binomial data, item difficulty is the same across items for an individual, while for the compound binomial model, this is not true. (This will be discussed further in a later section.) A program called DATGEN, developed by Hambleton and Rovinelli (1973), was used for the compound binomial case. This program, which is

usually used to simulate logistic test model data, produces a set of response patterns and simulated test scores to represent the performance of examinees. By varying certain of the parameters of the program, compound binomial test data can be generated.

#### Number of Examinees

For a small number of examinees, statistical indices generated by repeating a simulation are likely to differ greatly. For that reason, two hundred examinees were chosen as the number to use in this study. This is a sufficient number to generate stable estimates of reliability and validity.

#### Cut-off Scores

First, a domain score cut-off score must be specified. Based on previous research (Block, 1972), it was decided that .8 was a suitable value.

A cut-off score, called an advancement score, must also be set for the simulated test data, and this was varied with test length. An attempt was made to have the advancement score coincide as closely as possible, when expressed as a percentage of items, with .80. For instance, with a ten item test, the advancement score would be 8 items correct. However, for an eight item test, there is no exact number of items possible to form the correspondence. In this case, the advancement score was set as close as possible below (i.e., 6 items) and above (7 items), and both values were studied in the simulations.

Test Length

A number of test lengths, and associated cut-off scores, were specified. These lengths range from 2 to 40 items. The chart below presents the 17 tests lengths considered, along with advancement scores.

<u>No. of Items</u>	<u>Advancement Score</u>
2	1
2	2
4	3
4	4
6	4
6	5
8	6
8	7
10	7
10	8
10	9
15	11
15	12
15	13
20	16
20	18
40	32

Reliability and Validity Indices

A number of different indices were used to summarize reliability and validity information. For reliability, these are:

1. proportion agreement
2. coefficient kappa
3. Subkoviak's one administration estimate of proportion agreement

i. based on  $\hat{\pi}_i = \frac{X_i}{n}$

ii. based on  $\hat{\pi}_i = \hat{\pi}_i r + \pi_i(1-r)$

For validity, the indices include:

1. proportion agreement
2. a validity coefficient
3. four-fold utility
4. two-fold disutility
5. Livingston's unweighted efficiency
6. Livingston's weighted efficiency

#### Weightings of Classification Frequencies

In certain instances, an equal weighting of false-positive and false-negative classification errors, and also of correct classification frequencies, is questionable. Thus, both equal and unequal weights were considered. Unequal weights will influence the following three indices: four-fold utility, two-fold utility, and Livingston's weighted efficiency index. Using the following contingency tables, the weights used in the simulation are presented in the respective cells.

#### Equal Weights

		True Status	
		M	NM
Test Score	M	.5	.5
	NM	.5	.5

Unequal Weights

		True Status	
		M	NM
Test Score	M	.5	-.750
	NM	-.375	.25

3.5.2 Simulation ProceduresBinomial Data

Using either the procedure for specifying a beta prior distribution or a user-supplied prior domain score distribution, a prior distribution is built. An examinee is then sampled from this distribution, and if his/her domain score is greater than or equal to .80, a value of 1 (to signify true master) is assigned. If the value is less than .80, a value of 0 (to signify true non-master) is assigned. Next, this domain score is used to simulate test performance on the test length in question. As an example, consider a four item test, and suppose the individual's domain score is .70. Now the examinee's performance on the four items is simulated. This is done by generating four random numbers from a uniform distribution on the interval (0, 1) and observing where these numbers lie in relationship to .70.

If the value is less than or equal to .70, the individual is considered to have passed the item, greater than .70 to have failed. These ones and zeros are totaled across the four items, and a score

from 0 to 4 obtained. This is then compared to the advancement score, which would be 3 or 4 depending on the particular simulation. The examinee then receives an overall score of 1 if his/her score is greater than or equal to three, and 0 otherwise. This simulation procedure is repeated for each examinee, thereby supplying the parallel forms data necessary for reliability index computation. Finally, this procedure is repeated for all 200 examinees sampled. Each examinee will then have a set of three numbers, either zeros or ones. A comparison of an examinee's domain score to the cut-off score generates a score of 0 or 1; likewise for the two simulated tests, except that the two scores for the examinee are compared to the advancement score.

The 0's and 1's for the 200 individuals are sufficient data to fill the cells of two fold contingency tables from which the reliability and validity indices being considered can be computed. Finally, this simulation procedure was repeated for the 17 test length advanced score combinations under consideration.

#### Compound Binomial Data

The simulation procedures for the compound binomial data are similar to that of the binomial case. The difference in the two situations is that for the compound binomial case, the items are allowed to vary in difficulty.

To utilize DATGEN, the user reads in specifications for the distribution of item difficulty, discrimination, and guessing parameters and ability parameters. Item difficulties can either

be sampled from a normal or rectangular distribution; likewise for the discrimination and guessing parameters. Ability can also be specified as normal or rectangular. The end result is a set of response patterns and test scores to represent the performance of examinees on binary-scored items based on the compound binomial distribution.

In applying DATGEN to the test length-reliability, validity relationship under study here, a number of variables were fixed and others altered. These include:

1. For logistic test data, in particular, the three parameter logistic model, a guessing parameter is included. However, in criterion-referenced testing situations, it makes little sense to simulate data that contains a guessing component. In typical criterion-referenced testing situations, the testing follows instruction on content, and guessing is minimal. Thus, the guessing parameter was set at 0.
2. While a discrimination parameter is often used with logistic test data, it is not necessary to specify this as a variable for the data being simulated in this study. Thus, the discrimination parameters were set at a typical value of .59 for all test lengths and simulations considered.
3. The critical parameters to be specified in DATGEN for this simulation were ability and difficulty. They were specified as:

Distribution	Difficulty	Ability
8	Rectangular [-1,1]	Normal (1,1)
9	Rectangular [-1,0]	Normal (1,1)
10	Rectangular [-1,1]	Normal (0,1)

The simulation of the item response data using DATGEN differs somewhat from the procedure for the binomial model. However, once

the data is simulated, the procedures exactly coincide. To utilize DATGEN, the number of examinees, shape and characteristics of the ability distribution, number of test items and characteristics of item parameters need to be specified. In this simulation, because the guessing parameter was set at 0 and the discrimination parameter at .59, the three-parameter model is reduced to a one-parameter model. A cut-off score was set on the ability scale to separate examinees into mastery states. The probability of an examinee  $i$  answering item  $j$  correctly (denoted  $p_{ij}$ ) is given by:

$$p_{ij} = \frac{e^{\theta - b_j}}{1 + e^{\theta - b_j}}$$

Next,  $p_{ij}$  is compared with a random number from a uniform distribution on the interval (0,1). If the random number is less than or equal to  $p_{ij}$ , a value of 1 is assigned, otherwise a zero. The remainder of the procedure coincides exactly with the description for the binomial model, and thus, will not be repeated here.

### 3.6 Results and Discussion

Data were simulated for ten typical criterion-referenced testing situations and the resulting reliability and validity indices are presented in Tables 3.6.2 to 3.6.11 which follow. More extensive tables, which include all seventeen simulated test lengths, are presented in Appendix One. For four of the ten simulations, binomial data were simulated through the specification of a user prior distribution, and for three of the ten, binomial data



Table 3.6.1

## Simulated Distributions Considered in the Study

Distribution	Generated From/ Type of Data	Equal Class. Freq. Used	Equal and Unequal Freq. Used	Skewedness	Peakedness in Relation to Cut-off or Relationship of Mean to Cut-off
1	User Prior/ Binomial		✓	High Negative	Highly peaked at slightly above cut-off
2	User Prior/ Binomial		✓	Negative	Peaked just below the cut-off
3	User Prior/ Binomial	✓		Negative	Moderately peaked at the cut-off
4	User Prior/ Binomial	✓		Extreme Negative	Highly peaked above the cut-off
5	Beta Prior/ Binomial		✓	High Negative	Highly peaked in vicinity of cut-off
6	Beta Prior/ Binomial		✓	Negative	Peaked near, but below, the cut-off
7	Beta Prior/ Binomial	✓		Little, if any skewness	Peaked below the cut-off
8	DATGEN/Com- pound Binomial	✓		Moderate Negative	Test score means at or close to cut-off
9	DATGEN/Com- pound Binomial	✓		Negative	Test score means near cut-off
10	DATGEN/Com- pound Binomial	✓		None	Test score means distant from the cut-off

were simulated through the specification of a beta prior. The remaining three simulations were performed utilizing DATGEN, which simulates compound binomial data. In addition, for certain of the simulations, both equal and unequal weighting of classification frequencies were utilized. Table 3.6.1 organizes the simulations by characteristics, also describing skewness, peakedness, and relationship of the simulated data to the cut-off of .80.

A number of general comments on the procedures used are appropriate to mention at this point:

1. For the majority of simulations, the domain score distributions were centered at or near the .80 cut-off score. Two examples of exceptions to this procedure include simulation seven, which is centered at a point considerably lower than .80, and simulation four, which is centered at an extremely high domain score level. Centering of distributions near the cut-off score was done for two reasons. One, such a centering patterns real test data from instructional settings where criterion-referenced tests are utilized. Tests are usually given following instruction in a content domain, and one should expect a distribution of the examinees' scores, assuming instruction was effective, to be peaked at the upper end of the distribution, somewhere near the true cut-off. Two, locating the domain score distribution near the cut-off score will insure that conservative estimates of reliability and validity will be obtained.
2. For certain of the test lengths being simulated (2, 4, 6, and 8 items), it is impossible to set an advancement score that exactly coincides with a cut-off of .80. For instance, for a six item test should the cut-off be 4 items (or .67) or 5 (.83)? For the simulated tests that had larger numbers of items, the lengths were chosen (10, 15, 20, and 32 respectively) so that exact cut-offs corresponding to .80 (8, 12, 16, 32 respectively) could be chosen. For the shorter tests, data was simulated for more than one advancement score, but the results reported in Tables 3.6.2-3.6.11 are only for the cut-off closest to .80. Appendix One contains the simulated data for all the cut-points. When forced to choose an advancement score that doesn't coincide with .80, problems occur when making comparisons across test lengths. In the ideal situation, where all advancement scores coincide with the cut-off,

the reliability and validity indices should increase for increasing test lengths, except for disutility, which should decrease. For certain of the sets of simulated data, this is not the case. Due to the problem just discussed, the indices do not steadily increase, and instead "flip-flops" in the reported results occur. Only with the longer test lengths (i.e., 10 and greater) does the increasing pattern become evident.

3. For certain of the simulations, even when the problem just discussed isn't evident, the reliability and validity indices presented don't demonstrate the expected increasing pattern. This is best demonstrated in the tables for the compound binomial data presented in Appendix One. For these simulations, the test means are also included. For instance, for distribution eight, the mean for a 15 item test with a cut-off of 12 was 10.10; for the 15 item test with a cut-off of 13 it was 11.13. In order for the indices to demonstrate the expected increasing trends, means such as the above should closely coincide. This, however, is not the case with randomly generated data, and should be understood when viewing the patterns in the indices over test lengths.
4. For certain of the distributions, the effects of an equal and unequal weighting of classification frequencies are presented for comparison. The following contingency tables review the equal and unequal weightings used:

		EQUAL WEIGHTING		UNEQUAL WEIGHTING	
		TM	TNM	TM	TNM
M		.5	-.5	.5	-.75
NM		-.5	.5	-.375	.25

(M = master, NM = non-master, TM = true master, TNM = true non-master.) The unequal weighting were decided upon in a somewhat arbitrary fashion, but dictated by the following two general considerations. One, it is usually more costly to commit a false-positive error (a true non-master classified as a master) than to commit a false-negative error (a true master classified as a non-master). Two, it is more beneficial to make a correct decision about mastery than non-mastery. The weights used reflect these two concerns.

5. After presenting the data, the selection of an "optimal" test length would be useful. Unfortunately, there are few, if any, guidelines to assist in making the choice. A somewhat arbitrary decision was made by this author to consider proportion agreement indices of approximately .75 as sufficient for choosing a test length. This means that for the test length chosen that either: (1) 75% of the time individuals are consistently classified upon retest, or (2) 75% of the time individuals are classified, based upon test results, into their true mastery status. Thus, test lengths will be selected where both the reliability and validity proportion agreement indices are upwards of .75.

Given the above comments, results from the ten individual simulations will now be presented.

#### Simulation One

This simulation, which is presented in Figure 3.6.1 with related indices in Table 3.6.2, involves a domain distribution that is peaked just above the .80 cut-off. There are a substantial number of domain scores below the .80 cut-off (44%). The resulting indices, based upon simulated binomial data, show the general increasing trend with the characteristic problem of inversions in the increasing trend at 4, 6, and 8 items for the reliability indices. An unexpected result, and probably due to the random nature of simulation procedures, is the lower reliability for the 15 item test compared to the 12 item. Using the arbitrary 75% proportion agreement figure, a 10 item test would be sufficiently reliable and an 8 item test sufficiently valid.

Interval	Percentages			
	Simulation One	Simulation Two	Simulation Three	Simulation Four
0 - .10	0	0	1	0
.11 - .20	0	1	1	0
.21 - .30	1	1	4	0
.31 - .40	1	4	6	0
.41 - .50	4	8	8	0
.51 - .60	8	12	10	0
.61 - .70	12	22	20	5
.71 - .80	18	32	22	25
.81 - .90	36	12	18	40
.91 - 1.00	20	8	10	30

Figure 3.6.1. Percentage distributions for simulations one thru four.

Table 3.6.2  
 Selected Test Lengths and Associated Cut-Offs for Simulation One:  
 Equal and Unequal Weights

No. of Items	Reliability				Validity								
	Cut-off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Unweighted Utility	Unweighted Disutility	Unweighted Efficiency	Weighted Utility	Weighted Disutility	Weighted Efficiency
2	2	.610	.154	.885	.565	.690	.364	.190	.155	.507	.109	.180	.548
4	3	.755	.396	.799	.692	.720	.445	.220	.140	.549	.146	.171	.466
6	5	.715	.402	.753	.588	.750	.489	.265	.118	.641	.153	.150	.624
8	6	.700	.286	.792	.707	.755	.509	.255	.123	.717	.175	.150	.664
10	8	.745	.442	.780	.712	.805	.607	.305	.098	.768	.239	.094	.806
15	12	.710	.399	.801	.735	.825	.644	.325	.088	.872	.225	.101	.838
20	16	.775	.542	.806	.766	.865	.726	.340	.080	.880	.264	.075	.871
40	32	.790	.570	.835	.813	.860	.715	.330	.085	.887	.258	.081	.871

### Simulation Two

This simulation, which is presented in Figure 3.6.1 with related indices in Table 3.6.3, involves a domain distribution that is peaked just below .80, and is a somewhat "flatter" distribution than distribution one. It is also not as negatively skewed, since 80% of the domain scores are below the .80 cut-off. The resulting indices, based on binomial data, show the expected increasing trends with some inversions in the validity indices for the shorter tests. The reliability and validity indices do not increase as rapidly as for distribution one, and a 15 item test fulfills the 75% proportion agreement for reliability and domain validity.

### Simulation Three

This simulation, which is presented in Figure 3.6.1 with related indices in Table 3.6.4, involves a domain distribution that is peaked, but not highly, near .80. The indices, which again are based on simulated binomial data, show the general increasing trend with some indices for the 15 item test appear questionable. For this distribution, a 10 item test satisfies the 75% selection figure.

### Simulation Four

This simulation, which is presented in Figure 3.6.1 with related indices in Table 3.6.5, involves a domain distribution that is highly negatively skewed and peaked above .80 so that only 40% of the examinees' domain scores fall below .80. This would be characteristic of a criterion-referenced test given after highly effective instruction

Table 3.6.3

Selected Test Lengths and Associated Cut-Offs for Simulation Two:  
Equal and Unequal Weights

No. of Items	Reliability				Validity								
	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Unweighted Utility	Unweighted Disutility	Unweighted Efficiency	Weighted Utility	Weighted Disutility	Weighted Efficiency
2	2	.535	.073	.842	.527	.600	.280	.135	.183	.386	-.095	.285	.368
4	3	.660	.258	.769	.621	.515	.266	.040	.230	.246	-.183	.356	.253
6	5	.695	.381	.755	.665	.715	.430	.235	.133	.639	.019	.203	.629
8	6	.700	.398	.752	.681	.650	.405	.150	.175	.595	-.048	.257	.597
10	8	.735	.463	.754	.690	.690	.416	.190	.155	.663	.026	.203	.662
15	12	.745	.469	.772	.712	.755	.531	.255	.123	.784	.108	.137	.791
20	16	.755	.476	.807	.755	.830	.641	.330	.085	.850	.117	.133	.811
40	32	.830	.584	.828	.836	.875	.663	.375	.063	.945	.201	.071	.942



Table 3.6.4  
 Selected Test Lengths and Associated Cut-Offs for Simulation Three:  
 Equal Weights

No. of Items	Reliability				Validity					
	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Efficiency
2	2	.585	.169	.869	.580	.660	.376	.160	.170	.555
4	3	.695	.375	.760	.615	.665	.435	.165	.168	.544
6	5	.725	.440	.760	.670	.725	.459	.225	.138	.733
8	6	.685	.370	.766	.691	.720	.500	.220	.140	.675
10	8	.760	.514	.776	.721	.775	.562	.275	.113	.799
15	12	.735	.450	.804	.757	.820	.613	.320	.090	.851
20	16	.800	.580	.813	.793	.875	.736	.375	.063	.918
40	32	.830	.616	.855	.842	.890	.756	.390	.055	.928

Table 3.6.5  
 Selected Test Lengths and Associated Cut-Offs for Simulation Four:  
 Equal Weights

No. of Items	Reliability				Validity					
	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Efficiency
2	2	.600	.030	.900	.576	.685	.232	.185	.158	.520
4	3	.770	.128	.816	.817	.715	.214	.215	.143	.574
6	5	.695	.149	.774	.732	.720	.308	.220	.140	.652
8	6	.725	.170	.807	.786	.775	.416	.275	.113	.686
10	8	.735	.279	.794	.721	.765	.409	.265	.118	.702
15	12	.760	.372	.787	.723	.810	.532	.310	.095	.790
20	16	.770	.403	.784	.741	.815	.532	.315	.093	.797
40	32	.785	.486	.817	.791	.865	.677	.365	.068	.872

<u>Distribution Five</u>		<u>Distribution Six</u>	
<u>Quartile</u>	<u>Percentage</u>	<u>Quartile</u>	<u>Percentage</u>
25	70	25	50
50	80	50	80
75	90	75	95

<u>Distribution Seven</u>	
<u>Quartile</u>	<u>Percentage</u>
25	40
50	50
75	60

<u>Distributions Eight, Nine and Ten</u>		
<u>Distribution</u>	<u>Difficulty</u>	<u>Ability</u>
8	Rectangular [-1, 1]	Normal $\phi(1,1)$
9	Rectangular [-1, 0]	Normal $\phi(1,1)$
10	Rectangular [-1, 1]	Normal $\phi(0,1)$

Figure 3.6.2. Quartiles and difficulty and ability parameters for simulations five thru ten.

Table 3.6.6

Selected Test Lengths and Associated Cut-Offs for Simulation Five:  
Equal and Unequal Weights

No. of Items	Reliability				Validity								
	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Unweighted Utility	Unweighted Disutility	Unweighted Efficiency	Weighted Utility	Weighted Disutility	Weighted Efficiency
2	2	.550	.049	.874	.523	.660	.325	.200	.150	.479	.051	.208	.439
4	3	.690	.127	.785	.699	.640	.347	.140	.180	.385	.016	.253	.182
6	5	.695	.338	.777	.636	.700	.433	.200	.150	.493	.107	.178	.634
8	6	.715	.289	.786	.710	.740	.532	.240	.130	.554	.136	.165	.667
10	8	.735	.444	.761	.658	.745	.501	.245	.128	.670	.123	.165	.672
15	12	.745	.479	.799	.758	.770	.604	.245	.128	.692	.158	.143	.726
20	16	.760	.509	.817	.774	.835	.670	.335	.083	.839	.215	.113	.795
40	32	.805	.608	.835	.803	.855	.713	.365	.068	.875	.234	.090	.863

Table 3.6.7

Selected Test Lengths and Associated Cut-Offs for Simulation Six:  
Equal and Unequal Weights

No. of Items	Reliability				Validity								
	Cut-off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Unweighted Utility	Unweighted Disutility	Unweighted Efficiency	Weighted Utility	Weighted Disutility	Weighted Efficiency
2	2	.720	.439	.867	.625	.760	.540	.330	.085	.670	.136	.158	.645
4	3	.740	.433	.811	.720	.765	.563	.265	.118	.649	.085	.195	.544
6	5	.755	.491	.817	.762	.795	.596	.295	.103	.766	.188	.131	.760
8	6	.795	.580	.837	.794	.805	.649	.305	.098	.769	.217	.118	.742
10	8	.810	.618	.846	.807	.870	.754	.370	.065	.893	.241	.084	.884
15	12	.865	.724	.862	.838	.870	.746	.400	.050	.898	.247	.084	.891
20	16	.870	.740	.870	.868	.885	.773	.405	.048	.940	.259	.073	.945
40	32	.920	.839	.923	.918	.935	.870	.435	.033	.976	.336	.021	.991

with instruction. While the indices are high, the failure rate is also high. For instance, for the 10 item test, 40% of the examinees were consistent non-masters for the two simulated testing situations.

#### Simulation Seven

This simulation (Figure 3.6.2, Table 3.6.8) involves a distribution that represents a situation where the test is too difficult for the population (domain scores are centered at or near .50). This could occur when a test is given without instruction on a content domain. For example, for the 10 item test, the failure rate is 80% across the two administrations. The reliability and validity indices are very high for the longer tests demonstrating that, for this distribution, greater than 10 items on a test adds little to the consistency or accuracy of decision-making. The inversions evident for the shorter length tests, because of the differing advancement scores make it difficult to select a test length to satisfy the 75% selection figure. What is obvious is that a short test in this situation will lead to reliable and valid assessments. Because there are few examinees near the cut-off, the chances of misclassification errors are low, and hence, shorter tests lead to consistent and valid instructional decisions.

The next three simulations, utilized DATGEN, and thus simulated compound binomial data. Rather than sampling from a pre-specified distribution, as was done for simulations one thru seven, a different procedure was used. As a result, a test score distribution was formed

Table 3.6.8

Selected Test Lengths and Associated Cut-Offs for Simulation Seven:  
Equal Classification Weights

No. of Items	Cut-Off	Reliability				Validity				
		Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Unweighted Utility	Unweighted Disutility	Unweighted Efficiency
2	2	.610	.080	.829	.620	.710	.154	.210	.145	.550
4	3	.640	.210	.726	.576	.645	.133	.145	.178	.456
6	5	.620	.220	.802	.795	.845	.163	.345	.078	.776
8	6	.710	.173	.777	.724	.810	.147	.310	.095	.727
10	8	.840	.183	.815	.853	.895	.220	.395	.053	.877
15	12	.910	.352	.893	.899	.935	.469	.435	.033	.955
20	16	.960	.536	.910	.947	.950	.494	.450	.025	.969
40	32	.985	.524	.942	.961	.970	.568	.470	.015	.982

for each test length. Rather than discussing all these test score distributions, a ten item test was chosen as an appropriate example. Such a distribution will be presented with each of the following three simulations.

#### Simulation Eight

This simulation (Figure 3.6.2, Table 3.6.9), which was produced utilizing DATGEN, and thus is based upon compound binomial data, represents an interaction of moderate to difficult test items with a bright group. The resulting test score distributions are flat and negatively skewed with a mean at about 67%. For instance, for a 10 item test, the test score distribution is as follows:

<u>Test Score</u>	<u>Frequency</u>
10	26
9	32
8	41
7	37
6	22
5	17
4	14
3	4
2	5
1	2

(These test score distributions were produced only for compound binomial data.)

Because a number of examinee scores are near the cut-off, there is a possibility of misclassification errors and a longer test is necessary. An 8 item test satisfies the 75% figure for reliability and a 10 item test for validity.



Simulation Nine

This simulation (Figure 3.6.2, Table 3.6.10) represents an interaction of a moderate to easy test with a bright group. The test score distribution for a 10 item test is:

<u>Test Score</u>	<u>Frequency</u>
10	43
9	48
8	33
7	24
6	20
5	16
4	7
3	6
2	2
1	1

The resulting reliability and validity indices are higher for shorter test lengths than they were for simulation eight. An 8 item test satisfies the 75% proportion agreement selection point.

Simulation Ten

This simulation (Figure 3.6.2, Table 3.6.11) represents an interaction of moderate to difficult test items with an average group. The test score distribution for the 10 item test being used as an example is:

<u>Test Score</u>	<u>Frequency</u>
10	1
9	9
8	22
7	24
6	26
5	25
4	28
3	27
2	23
1	9
0	6

Table 3.6.9  
Selected Test Lengths and Associated Cut-Offs for Simulation Eight:  
Equal Weights

No. of Items	Reliability				Validity					
	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Efficiency
2	2	.615	.234	.827	.509	.650	.338	.150	.175	.394
4	3	.740	.412	.783	.657	.720	.522	.220	.140	.559
6	5	.650	.287	.737	.576	.720	.460	.220	.140	.615
8	6	.745	.469	.788	.686	.725	.507	.225	.138	.670
10	8	.715	.430	.761	.677	.765	.548	.265	.118	.744
15	12	.725	.435	.800	.739	.820	.630	.320	.090	.845
20	16	.745	.470	.813	.768	.850	.688	.350	.075	.891
40	32	.840	.659	.842	.821	.880	.740	.380	.060	.939

Table 3.6.10  
 Selected Test Lengths and Associated Cut-Offs for Simulation Nine:  
 Equal Weights

No. of Items	Reliability				Validity					
	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Efficiency
2	2	.700	.372	.882	.577	.715	.458	.215	.143	.582
4	3	.655	.086	.791	.673	.665	.387	.165	.168	.413
6	5	.730	.429	.771	.634	.740	.499	.240	.130	.653
8	6	.790	.378	.811	.765	.750	.444	.250	.125	.607
10	8	.740	.437	.789	.711	.800	.587	.300	.100	.753
15	12	.785	.551	.795	.755	.810	.628	.310	.095	.792
20	16	.805	.602	.810	.742	.800	.599	.300	.100	.821
40	32	.815	.610	.837	.808	.855	.715	.355	.073	.884

Table 3.6.11  
 Selected Test Lengths and Associated Cut-Offs for Simulation Ten:  
 Equal Weights

No. of Items	Reliability				Validity					
	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Efficiency
2	2	.705	.230	.837	.713	.805	.193	.305	.098	.782
4	3	.680	.312	.735	.584	.650	.227	.150	.175	.536
6	5	.865	.413	.863	.858	.900	.482	.400	.050	.900
8	6	.805	.532	.801	.737	.735	.353	.235	.133	.771
10	8	.860	.466	.872	.866	.900	.546	.400	.050	.929
15	12	.920	.294	.892	.913	.940	.441	.440	.030	.970
20	16	.930	.475	.902	.887	.945	.450	.445	.027	.975
40	32	.945	.672	.929	.937	.955	.713	.455	.023	.985

This distribution demonstrates the effect that a cut-score which is widely disparate from the mean of a distribution has on reliability and validity. The test score distribution indicates that the scores for this simulation are centered at .50 (or 50% of the items), and that there is little or no skew (-.03). The distribution is essentially uniform in the interval from .20 to .80. With a .80 cut-off, there will be virtually no misclassifications and a short test will lead to reliable and valid assessment. Such a distribution is not likely in criterion-referenced instructional situations, but is more likely the result of a difficult norm-referenced test given to moderately able students.

Next, a number of comments will be made that pertain to the ten simulations discussed:

1. The three simulations that involved compound binomial data resulted in a similar test lengths as those for binomial data in reaching acceptable levels of reliability. (Subkoviak [1976] reported a similar finding.) For instance, simulation eight, which involves compound binomial data, is similar to simulation three involving binomial data. Both simulations required tests of around 10 items to satisfy the 75% selection criterion. In a like fashion, simulation ten and simulation seven can be compared. Both of these simulations involved distributions that were symmetrical and with average test scores equal to about 50%. For both simulations, very short tests led to reliable and valid decision-making due to the likelihood of few, if any, misclassification errors.
2. Two estimates of proportion agreement as an indication of test-retest reliability were utilized in this study; both estimates are based upon data collected from a single test. The first estimate, labeled Subkoviak (1) in the Tables, makes use of proportion correct ( $\hat{\pi}_i$ ) in estimating proportion agreement, while the second ( $\hat{\hat{\pi}}_i$ ), labeled Subkoviak (2) makes use of collateral group information. In comparing these two estimates to the actual proportion agreement index obtained, it was decided that a "closest to" criterion would be employed to try to draw some overall conclusions

about which index functioned the best as an estimate. Thus, in Table 3.6.12, if, for instance, for simulation four, with a test length of four items, the Subkoviak (2) estimate was closest, it was scored a one. This was done separately for the binomial and compound binomial data, across all ten simulations.

Based upon Table 3.6.12 just presented and Tables 3.6.2-3.6.11, the following comments can be made concerning the binomial data. For the short test lengths, neither estimate was very accurate, nor was there a clear pattern of dominance utilizing the "closest to" criterion (except for lengths of 2 items). For the middle test lengths, the Subkoviak (2) estimate performed better, and this was true also for longer test lengths, except that now both estimates tended to be quite accurate. The above pattern did not appear with the compound binomial data, and this is why this data was presented separately. However, because of the small number of simulations, no attempt to interpret patterns will be made for the compound binomial data.

3. A comparison of those indices effected by equal and unequal weightings of classification frequencies demonstrated that for all distributions where both equal and unequal weights were used, unweighted utility was higher than weighted utility, and vice-versa for disutility. This, of course, is a direct result of the weights used. The patterns do, however, coincide with expectations based upon the weights assigned. A common trend in the comparison of weighted and unweighted efficiency is not so apparent. The indices fairly closely coincide, particularly for the longer tests. For some simulations (i.e., simulation six), all values of unweighted efficiency are higher than weighted efficiency, while for other distributions (e.g., five) there is a high level of "flip-flopping" and no discernible trend is apparent. What is apparent is that longer tests are more efficient, but that like the other indices, there is a point where increasing the length of the test does little to enhance the test's reliability or validity.
4. The results of setting differing advancement scores for tests of the same overall length resulted in indices that could be predicted in a comparative fashion. For distributions peaked above the cut-off, the effects of increasing the advancement score was to decrease the proportion agreement reliability and validity indices. For distributions peaked below the cut-off, the opposite was true; as advancement scores increased, so did the proportion agreement reliability and validity indices. Finally, for

Table 3.6.12

A Comparison of Subkoviak (1) and Subkoviak (2)  
Estimates of Proportion Agreement

Binomial Data			
Test Length	Subkoviak (1)	Subkoviak (2)	"Tie"
2	0	7	
4	4	3	
6	2	5	
8	1	6	
10	2	5	
15	1	4	2
20	1	4	2
40	0	5	2

Compound Binomial Data			
Test Length	Subkoviak (1)	Subkoviak (2)	"Tie"
2	0	2	1
4	2	1	
6	2	0	1
8	3	0	
10	0	1	2
15	1	2	
20	2	1	
40	1	2	

distributions peaked near .8, the proportion agreement reliability and validity indices changed in the expected direction, but much less rapidly. Table 3.6.13 summarizes these results for three simulations for test lengths of 10 items, with advancement scores of 7, 8, and 9 items. Appendix One contains a complete listing of the relevant data.

5. In the seven binomial data simulations, the proportion agreement indices for reliability and validity increased from a low around .55 for a two item test to .85 for a forty item test. For the three compound binomial simulations, the increasing pattern was again evident, and the results were similar to those involving the binomial data. These values varied across simulations depending upon the peakedness of the distribution and where the distribution was centered in reference to .80, the cut-off score. For most of the simulations, suitable reliability and validity indices were generated for tests of around 8 or 10 items.

### 3.7 Conclusions

At the beginning of this chapter, three objectives were specified to guide the research involving the relationship of test length to criterion-referenced reliability and validity.

In reference to the first objective, a computer program was developed with the assistance of Frederick DeFriesse. The completion of objectives two and three were accomplished through the research presented in this chapter. Ten different simulations were performed, each involving seventeen different test lengths, and tables relating test lengths to indices of reliability and validity were also presented. The ten simulations were designed to cover the range of testing situations most often encountered when utilizing criterion-referenced tests.



Table 3.6.13

The Relationship of Cut-off Scores to the Simulated  
Distribution for Three Selected Distributions

Distribution	Characteristic	Length	Advance- ment Score	Reliability	Validity
4	Peaked above .80	10	7	.840	.795
		10	8	.735	.765
		10	9	.615	.710
7	Peaked below .80	10	7	.720	.765
		10	8	.840	.895
		10	9	.935	.945
5	Peaked just below .80	10	7	.735	.670
		10	8	.745	.740
		10	9	.765	.765

To summarize the research presented in this chapter, the following comments can be made. When little empirical research has been done in an area, a simulation study is often a reasonable starting point. The simulated data relating test length to reliability and validity indices reported in this dissertation will hopefully serve a two-fold purpose. One, the research reported here should lead to empirical real-data investigations of test lengths to reliability and validity. Two, until such research is undertaken, the indices reported here can provide the practitioner working in the area with needed estimates of reliability and validity.

C H A P T E R   I V  
SETTING STANDARDS FOR CRITERION-REFERENCED TESTS AND  
AN APPLICATION TO MINIMUM COMPETENCY TESTING

4.1 Introduction

Hambleton, Swaminathan, Algina, and Coulson (1978) have discussed two major uses for test scores derived from criterion-referenced tests: domain score estimation and allocation of examinees to mastery states. The second use, the allocation of examinees to mastery states, depends on the existence of a well-defined performance standard, or cut-off score. (This chapter will follow the work of other authors in this area and use the terms performance standard, proficiency standard, cut-off score, and passing score interchangeably.) The focus is on how much an individual knows in reference to a well-defined subject domain and a specified standard of performance. Based upon an individual's score on a test, where the test items serve as a representative sample from a subject domain, a mastery decision is made.

Thus, it can be seen that in a criterion-referenced testing situation, a cut-off score (there can be several cut-off scores, although usually only one is set) must be set, based upon a unit or domain of study, in order to make a decision about an individual's mastery status. The results of this decision will depend upon the context within which the test is being used. As an example,

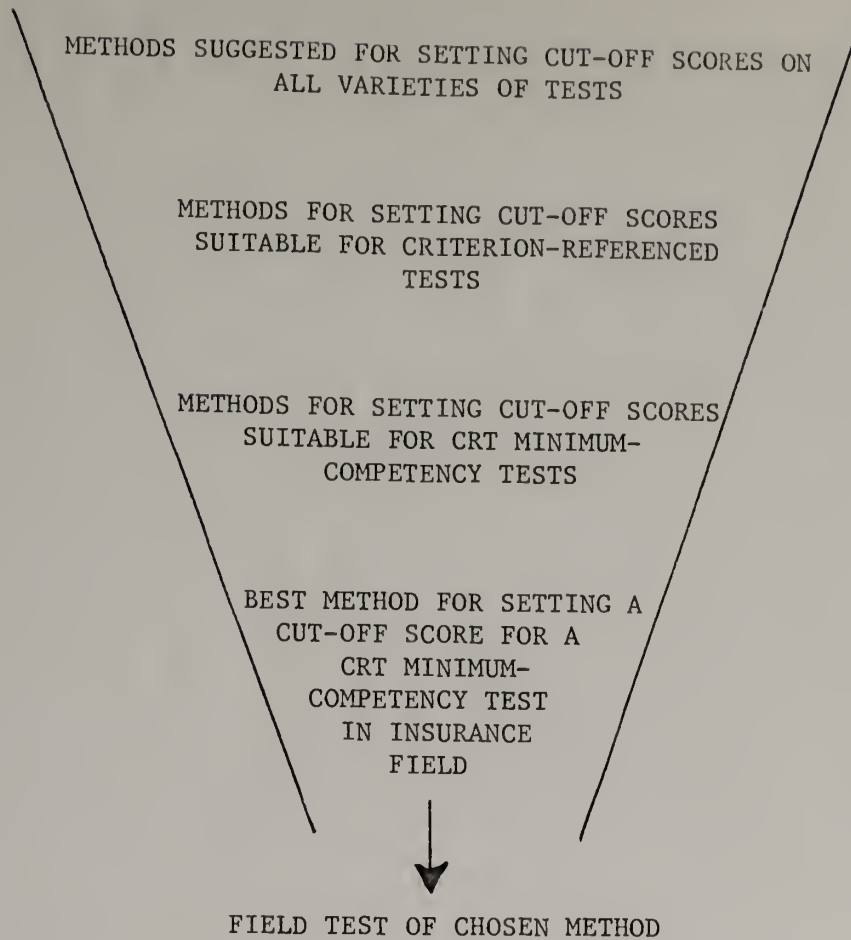
consider the Mastery Learning paradigm (Bloom, 1968; Block, 1972). In this situation, if a student's score exceeds the cutting score, he/she is advanced to the next unit of instruction. If the student's score falls below the standard, remedial activities are prescribed. It is important to understand that the decision being made is on the level of the individual, and as such, the status of other individuals should not enter into the decision. The passing score, or standard, should be set by a process that takes into account more information than simply how other individuals perform on the test.

Given what has just been said about the importance of cut-off scores for proper criterion-referenced test score usage, one would think that this would be a well-researched and documented area. This is simply not the case. Most of the work done to date has been concerned with the suggestion of possible methods, perhaps twenty-five in number, rather than with actual empirical investigation. In addition to the individual work done, there have been two excellent reviews of cut-score procedures advanced (Millman, 1973; Meskauskas, 1976), and one recent review that was highly critical of the field (Glass, 1978a).

This chapter will consist of three parts, and can be conceptualized as a "funneling process"; that is, a focusing in on a particular cut-score method to use to solve a particular test usage situation. The first section will be a review of the cut-score methods advanced to date. The review will draw on the work of Millman, Meskauskas and Glass, adding the many recently advanced

cut-score procedures. In this review, a number of suggested (and often utilized) procedures that are of minimal use for criterion-referenced testing situations will also be discussed.

The second section of the "funneling process" will be concerned with the application of criterion-referenced testing cut-score procedures to the minimum competency testing movement. As will be demonstrated, a number of the cut-score methods that are highly useful in other contexts are not useful for the determination of minimum competency standards. Reasons why many of the methods are not suitable will be offered as well as why the methods remaining are the sifting process are useful. A suggestion about which methods are most useful for minimum competency testing will also be offered. Finally, a field application of one of the suggested methods will be described in detail. This test involved the setting of a cut-score on a number of minimum competency certification tests in the insurance area. Practical implementation suggestions, based upon the field application, will also be presented. Pictorially, the material in this chapter can be represented as follows:



#### 4.2 Methods for Setting Cut-Off Scores Suitable for Criterion- Referenced Tests

In this section, a number of procedures that are not useful for setting cut-off scores for criterion-referenced tests will first be discussed. Then the remaining methods that are useful will be discussed in greater detail. Figure 4.2.1 is most useful as a starting point. In Figure 4.2.1, three sets of procedures for setting cut-off scores are presented. As the ensuing discussion will demonstrate, most of the procedures in Figure 4.2.1 are not useful for the setting of cut-off scores on criterion-referenced tests.

Traditional

Normative

Test  
Distribution

Performance  
By a  
Criterion  
Group

Models Based on  
Distributional Assumptions  
About Ability

State  
Models

Continuum  
Models

Emrick (1971)  
Roudabush (1974)  
Macready and  
Dayton (1977)

Figure 4.2.1 Three general sets of procedures for setting standards.

Traditional standards are standards that have gained acceptance simply because of their frequent use. Classroom examples include the 90 to 100 percent is an A, 80 to 89 percent is a B, etc. A larger scale example of a traditional standard is the cut-score point of 65 sets by the New York State Board of Regents for the regents exams administered in New York. Such standards should not be used for criterion-referenced tests (or for norm-referenced tests either) because they simply do not consider relevant information, such as the difficulty of the test, in setting the cut-point. New York, for instance, periodically has trouble with the 65 cut-off applied to the physics exams. While tests used from one year to the next are constructed to be parallel, they often vary somewhat in difficulty, and hence a 65 does not always mean the same thing for two tests. A student, assuming no learning or practice effect (i.e., constant knowledge), could fail one test and pass another. The Board of Regents often must adjust scores on the harder tests, but this can't completely alleviate the problem, and always questions are raised by teachers, students, and parents.

Another example of the use of traditional standards has been discussed in detail by Glass (1978a, 1978b). This concerns the more or less arbitrary setting of the cut-off points in reading and mathematics for the Florida Competency Testing Program at the traditional 70%. The results of this decision led to a 38 percent failure rate on the math test and a 10 percent failure rate on reading. Without considering the subject matter and the difficulty of the items comprising the test in setting the cut-off, it is



impossible for Florida to ascertain whether the difference is a matter of legitimate concern or simply caused by differences in test difficulty. Florida seems to be proceeding as though the differences are real, and hence is spending vast amounts for mathematics remediation (see Glass, 1978b). When the court cases involved with these tests begin to occur is simply a matter of legal processing time.

Normative standards actually could refer to three different uses of normative data, two of which are, at best, questionable. The first method makes use of the normative performance of some external "criterion" group. The most recent example of such a procedure has been cited by Jaeger (1978); it concerns the Adult Performance Level tests administered by Palm Beach, Florida schools. Test performance of groups of "successful" adults were used to set cut-off points for high school students. The notion was that these levels would be necessary for the high school student to succeed once out of school. Such a procedure can be criticized on a number of grounds. Jaeger (1978) points out that society changes, and that performance standards should also change. Hence, performance standards based on adult performance may simply not be relevant for high school students. Shepard (1976) has pointed out that any normatively-determined standard will immediately result in a multitude of counter-examples. Further, Burton (1978) recently pointed out that relationships between skills in school subjects and later success in life is not readily determinable, hence observing the degree of achievement of some "successful" norms group makes little sense. Jaeger (1978) states Burton's point as follows" "There

are no empirically tenable survival standards on school-based skills that can be justified through external means." While the above example concerns a minimum competency testing situation, the use of an external criterion group is questionable for all criterion-referenced testing situations. This is because of the difficulty involved in establishing the relationship between the criterion group and the group in question. Only in the simplest of situations, such as when the criterion-group is instructed or uninstructed (Berk, 1976), does this procedure begin to have relevance.

A second way of proceeding with normative data is to make a decision about a cut-off based on the distribution of scores of students who take the test. This avoids decisions about a criterion group, but it is still questionable for setting standards. For instance, Glass (1978a) cites the California High School Proficiency Examination, where the 50th percentile of graduating seniors constitutes the standard. Little can be said of a procedure where whether or not an individual passes or fails depends on the other people taking the test. In the California situation, the standard was set with no reference at all to the content of the test and the difficulty of the constituent items.

The third use of normative data discussed in the literature concerns the supplemental use of normative data in setting a standard. Researchers in this field, such as Jaeger (1978), Shepard (1976) and Conaway (1976, 1977) all favor such a procedure. Jaeger's (1978) recently advanced procedure for setting cut-off scores, to

be discussed later in this chapter, calls for incorporation of some tryout test data with judgmental data. Shepard (1976) makes the following point concerning normative data:

Expert judges ought to be provided with normative data in their deliberations. Instead of relying on their experience, which may have been with unusual students or professionals, experts ought to have access to representative norms. . . of course, the norms are not automatically the standards. Experts still have to decide what "ought" to be, but they can establish more reasonable expectations if they know what current performance is then if they deliberate in a vacuum. [p. 30]

This author agrees with Jaeger, Conaway, and Shepard about the usefulness of normative data used in conjunction with one of the standard setting methods to be discussed. Further, Jaeger (1978) and Shepard make the supplemental point that standard setting methods should also be iterative. If this is to be the case, then the normative data should be used to get the process initiated. At some point in time, the standard should stand "on its own" without the supplemental normative data. The normative data would then act as a catalyst in getting the standard setting method initiated.

In reference to models based on ability assumptions, Meskauskas (1976) has pointed out the differences between the continuum and state model conceptualizations of the ability being measured by the test. According to Meskauskas, two characteristics of continuum models are:

1. Mastery is viewed as a continuously distributed ability or set of abilities.
2. An area is identified at the upper level of this continuum, and if an individual equals or exceeds the lower bound of this area, he/she is termed a master.

State models, rather than being based on a continuum of mastery, view mastery as an all-or-none description of learning state.

Three characteristics of state models are:

1. Test (true) score performance is viewed as an all-or-nothing state.
2. The standard or cut-score is thus implicitly set at 100%.
3. A consideration of measurement error results in the setting of a cut-off score of less than 100%.

When viewing the issue of the measurement of a well-defined content domain, as is the case in criterion-referenced testing, a continuum model seems to offer by far the most potential. When the issue becomes the measurement of minimum competency utilizing a written criterion-referenced test, a continuum model conceptualization is essential. The ability scale must be treated as a continuum in order for the setting of a cut-point separating minimally competence from incompetence to take place.

There are at least three methods for setting cut-off scores that are built on a state model conceptualization of mastery. The models take into account measurement and other variables in "tempering" the standard from 100%. State model procedures advanced include Emrick's mastery testing evaluation model (1971), Roudabush's work on true-score models (1974), and more recently, the work of Macready and Dayton (1977). In sum, because of the assumptions made about mastery ability status, these models are not useful for criterion-referenced testing. The models appear, however, to hold great merit for performance testing, where the ability being measured is often either present or absent.

Thus far, the material presented in this section has concerned methods of setting cut-off scores that are not useful for criterion-referenced tests. What remains of Figure 4.2.1 that has not been discussed are the continuum models for assessing mastery. Figure 4.2.2 represents a breakdown of these models in conjunction with certain models where the underlying distributional assumption on ability has not been articulated. These models and methods are all suitable for criterion-referenced testing situations, and will now be discussed in greater detail. However, before doing so, one very important point must be made. At some point in the application of each method, judgmental data is involved. Jaeger (1976) makes this point when stating that all standard setting procedures are judgmental. Therefore, the trichotomy presented may be misleading without further clarification as to what is meant by judgmental models in Figure 4.2.2. The judgmental models organized together in Figure 4.2.2 are based upon the active involvement of a panel of judges who assess the individual items constituting the test, or decide on the presence of guessing (or item sampling error). While judgments are involved with the other procedures, which we will point out, these judgments are of a somewhat different variety. Thus, the trichotomy presented really only serves an organizational function; it is not technically accurate.

The models and methods will now be discussed in greater detail, starting first with judgmental models. The second group of models to be discussed are the empirical models, and finally those models

<u>Judgmental Models</u>		<u>Combination Models</u>		<u>Empirical Models</u>	
<u>Item Content</u>	<u>Guessing</u>	<u>Judgemental-Empirical</u>	<u>Data—Two Groups</u>	<u>Data-Criterion Measure</u>	
Nedelsky (1954)	Millman (1973)	Contrasting Groups (Zieky and Livingston, 1977)	Berk (1976)	Livingston (1975)	
Modified Nedelsky (Nssisif, 1978)		Borderline Group (Zieky and Livingston, 1977)		Livingston (1976)	
Angoff (1971)				Huynh (1976)	
Modified Angoff (ETS, 1976)				Huynh and Perney (1977)	
Ebel (1972)			<u>Educational Consequences</u>	Van der Linden and Mellenbergh (1977)	
Jaeger (1978)			Block (1972)		
		<u>Bayesian Methods</u>			
		Schoon, Gullion, Ferrara (1978)			
					<u>Decision-Theoretic<sup>1</sup></u>
					Kriewall (1972)

<sup>1</sup>In addition, there are a number of decision-theoretic models that deal with test length considerations. These are also applicable to cut-off score determination (see, for example, Millman, 1974).

Figure 4.2.2. A classification of models and methods for determining standards.

that are based on a combination of judgment and data will be discussed.

#### 4.2.1 Judgmental Models-Item Content

In this situation, individual items are inspected, with the level of concern being how the minimally competent person would perform on the items. In other words, a judge is asked to assess how or to what degree an individual who could be described as minimally competent would perform on each item. It should be noted before describing particular procedures utilizing this criterion that while this is a good deal more objective than setting standards based on any of the methods previously discussed, there is still a degree of subjectivity. The notion of minimal competence is arbitrary and subjective, and further, asking judges to assess it adds even more subjectivity. Bearing these concerns in mind, this author feels the procedures to be discussed have merit. Six procedures based on item content assessment will now be discussed.

##### i. Nedelsky's Method

In Nedelsky's method, judges are asked to view each question in a test with a particular criterion in mind. The criterion for each question is, which of the response options should the minimally competent student (Nedelsky calls them D-F students) be able to eliminate as incorrect. The minimum passing level (MPL) for that question then becomes the reciprocal of the remaining alternatives. For instance, if on a 5 alternative multiple choice question, a

judge feels that a minimally competent person could eliminate 2 of the options, then for that question,  $MPL = \frac{1}{3}$ . The judges proceed with each question in a like fashion, and upon completion of the judging process, sum the values for each question to get a minimum passing level on the total test ( $MPL_{FD}$ ). Finally, at this stage, the total-test MPL values are averaged across judges. The average is denoted  $\overline{MPL}_{FD}$ .

Nedelsky felt that if one were to compute the standard deviation of  $MPL_{FD}$ 's, that this distribution would be synonymous with the (hypothesized or theoretical) distribution of the scores of the borderline or D-F students. This standard deviation,  $\sigma$ , could then be multiplied by a constant K, decided upon by the test user, that would regulate how many (as a percent) of the borderline students pass or fail. The final formula then becomes:

$$MPL_F = \overline{MPL}_{FD} + K\sigma_{FD}$$

where the subscript F stands for final.

How does the  $K\sigma$  term work? Assuming an underlying normal distribution, if one sets  $K=1$ , then 84% of the borderline or D-F students will fail. If  $K=2$ , then 98% of these students will fail. If  $K=0$ , then 50% of the students on the borderline will fail. The value for K is set by the instructors prior to the examination.

The final result of the applications of Nedelsky's method should be an absolute standard for a cutting or minimum passing point. This is because the minimum passing point is arrived at in a manner independent of the score distributions of any reference



group. In fact, the standard is arrived at prior to application of the test to the group one is concerned about testing. However, while the standard might be called absolute, there is a great deal of subjectivity involved. While subjective ratings on the part of the judges or instructors are "washed out" somewhat by taking a mean, there is still a great deal of subjectivity that goes into determining the value of  $K$ . For this reason, and also based upon an article to be discussed later, we suggest that the method be used with caution.

The following somewhat simplistic example is included to demonstrate how the Nedelsky method can be applied in a criterion-referenced testing situation.

Example: Suppose 5 judges were asked to score, using the Nedelsky method, a 6 question criterion-referenced test made up of questions that have 5 response options each. Further, suppose the judges agreed that they would like 84% of the D-F or minimally competent students to fail (i.e., they set  $K = +1$ ). Given the following information, the minimum passing score is:

Judge	Question						Sum
	1	2	3	4	5	6	
A	.25*	.33	.25	.25	0	.33	1.41
B	.25	.50	.25	.50	.25	.33	2.08
C	.33	.33	.25	.33	.25	.33	1.82
D	.25	.33	.25	.33	.25	.33	1.74
E	0	.50	.25	.33	0	.25	1.33

\*The minimum passing level for the question. In a five option question, the possible values are 0, .25, .33, .50, and 1.0.

$$1. \text{MPL}_{\text{FD}} = \frac{\Sigma \text{MPL}_{\text{FD}}}{5} = \frac{8.38}{5} = 1.68$$

$$2. \sigma_{\text{FD}} = \sqrt{\frac{\Sigma (\text{MPL}_{\text{FD}} - \overline{\text{MPL}}_{\text{FD}})^2}{5}}$$

$\text{MPL}_{\text{FD}}$	$\text{MPL}_{\text{FD}} - \overline{\text{MPL}}_{\text{FD}}$	$(\text{MPL}_{\text{FD}} - \overline{\text{MPL}}_{\text{FD}})^2$
1.41	-.27	.073
2.08	.40	.160
1.82	.14	.020
1.74	.06	.004
1.33	-.35	.123

SUM = .380

$$\sigma_{\text{FD}} = \sqrt{\frac{.380}{5}} = \sqrt{.076} = .28$$

$$\begin{aligned} 3. \text{MPL}_{\text{FD}} &= \overline{\text{MPL}}_{\text{FD}} + K\sigma_{\text{FD}} \\ &= 1.68 + .28 \\ &= 1.96 \end{aligned}$$

Therefore, approximately two questions out of six is the minimum passing level on this test. From a practical standpoint, this value would seem low, but the data is created to demonstrate the process and not to model a real testing situation. Therefore, no practical significance should be attached to the answer.

#### ii. "Modified Nedelsky" Method

Nassif (1978), in setting standards for a competency-based teacher education and licensing system in the state of Georgia, utilized a "modified Nedelsky" approach to obtain a cut-off score.

Because of the number of items judges would have to assess, a simpler approach, where the item as a whole is rated, was decided upon. Each item was judged utilizing the question "Should a person with minimum competence in the teaching field be able to answer this item correctly?" Possible responses were "yes," "no," and "I don't know." The "yes" responses were then compared to binomial probability tables, and if the probability of receiving a given number of "yes" ratings was less than chance by one in ten, the item was classified as appropriate.

### iii. Ebel's Method

Ebel (1972) goes about arriving at a minimum passing score in a somewhat different manner from Nedelsky, yet his procedure is also based upon the test questions rather than an "outside" distribution of scores. Judges are asked to rate items along two dimensions, relevance and difficulty. Ebel uses four categories of relevance: Essential, important, acceptable and questionable, and three difficulty levels: Easy, medium and hard. These categories then form (in this case) a 3 x 4 grid. The judges are next asked to do two things:

1. Locate each of the test questions in the proper cell, based upon relevance and difficulty,
2. Assign a percentage to each cell; that percentage being the percentage of items in the cell that the minimally-qualified candidate should be able to answer.

Then the number of questions in each cell is multiplied by the appropriate percentage (agreed upon by the judges), and the sum of

all the cells, when divided by the total number of questions, yields the minimum passing score.

The example that follows is modeled after an example offered by Ebel (1972). Suppose that for a 100 item test, 5 judges came to the following agreement on percentage of success for the minimally qualified candidate.

Relevance	Difficulty Level		
	Easy	Medium	Hard
Essential	100%*	80%	--
Important	90%	70%	--
Acceptable	90%	40%	30%
Questionable	70%	50%	20%

\*The expected percentage of passing for items in the category.

Combining this data with the judge's location of test questions in the particular cells would yield a table like the following:

Item Category	Number of Items*	Expected Success	Number X Success
ESSENTIAL			
Easy	85	100	8500
Medium	55	80	4400
IMPORTANT			
Easy	123	90	11070
Medium	103	70	7210
ACCEPTABLE			
Easy	21	90	1890
Medium	43	40	1720
Hard	50	30	1500
QUESTIONABLE			
Easy	2	70	140
Medium	8	50	400
Hard	10	20	200
TOTAL	500		37030

$$\frac{37030}{500} = 74$$

\*The number of items placed in each category by all five of the judges.

The passing score would then be 74%.

Three comments can be made about Ebel's method that should be sufficient to convince people to be careful in using it. One, Ebel offers no prescription as to what the number or type of descriptions should be along the two dimensions. This is left up to the judgment of the individuals judging the items. It could likely be the case

that a different set of dimensions applied to the same test could yield a different passing score. Two, the process is based upon the decisions of judges, and while the standard might be called absolute in that it is referenced to no other distribution, it can't be called an "objective" standard. Three, a point about Ebel's method has been offered by Meskauskas (1976):

In Ebel's method, the judge must simulate the decision process of the examinee to obtain an accurate judgment and thus set an appropriate standard. Since the judge is more knowledgeable than the minimally-qualified individual, and since he is not forced to make a decision about each of the alternatives, it seems likely that the judge would tend to systematically over-simplify the examinee's task. . . . Even if this occurs only occasionally, it appears likely that, in contrast to the Nedelsky method, the Ebel method would allow the rater to ignore some of the finer discriminations that an examinee needs to make and would result in a standard that is more difficult to reach. [p. 138]

#### iv. Angoff's Method

Angoff's technique asks the judges to assign a probability to each question directly, thus circumventing the analysis of a grid or the analysis of response alternatives. Angoff (1971) states:

. . .ask each judge to state the *probability* that the "minimally acceptable person" would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score. [p. 515]

#### v. Modified Angoff Method

ETS (1976) has recently utilized a modification of Angoff's method for setting standards. Based on the rationale that the task of assigning probabilities may be overly difficult for the items to be assessed (items on the National Teachers Examinations), ETS instead supplied a seven point scale on which certain percentages were fixed. Judges are asked to estimate the percentage of minimally knowledgeable individuals who would know the answers to the question by selecting from the seven point scale:

5      20      40      60      75      90      95

ETS has also used scales with the fixed points at somewhat different values; the scales are consistent though in that seven points are given to choose from. The center point is chosen to coincide as closely as possible with the percent correct on past exams. Seventy is not used because it is a typical, or traditional, cut-point, and would probably be selected more often based solely on that fact.

#### vi. Jaeger's Method

Jaeger (1978) recently proposed a method for standard-setting on the North Carolina High School Competency Test. Jaeger's method incorporates a number of suggestions made by researchers in this field (Jaeger, 1976; Shepard, 1976; Conaway, 1976, 1977); it is iterative, based on judges from a variety of backgrounds, and employs normative data. Further, rather than asking a question

involving "minimal competence," a term which is hard to conceptualize and operationalize, Jaeger's questions are instead:

"Should every high school graduate be able to answer this item correctly?" "    Yes,     No."  
and "If a student *does not* answer this item correctly, should he/she be denied a high school diploma?" "    Yes,     No."

After a series of iterative processes involving judges from various areas of expertise, and after the presentation of some normative data, the passing scores determined by all groups of judges of the same type are pooled, and a median computed. Then the minimum median across all groups would be the pass-point.

#### vii. Studies Comparing Judgmental Methods

This author is aware of two studies that compare judgmental methods for setting cut-off scores; one study was done in 1976, the other is presently under way at ETS.

In 1976, Andrew and Hecht carried out a well publicized (see Glass, 1978a) empirical comparison of the Nedelsky and Ebel methods. In the study, judges met on two separate occasions to set standards for a 180 item, four options per item, exam to certify professional workers. On one occasion, the Nedelsky method was used, and the Ebel method in the other. The percentage of questions that should be answered correctly by the minimally competent person was 69% by the Ebel method and 46% by the Nedelsky method. While Glass (1978a) has chosen to make some distributional assumptions to accent further the differences between the methods, we prefer to refer to Meskauskas' comment presented earlier about how judges using Ebel's method might



tend to systematically oversimplify the examinee's task, and hence result in a higher standard.

Donald Rock at ETS is presently pursuing research on the use of the Nedelsky and Angoff methods for setting cutting scores on Real Estate Certification Examinations. The results of this study, which have not been released, should shed some light on the comparability of the two judgmental procedures probably used most frequently to date.

#### viii. Suggestions on Usage

All of these methods, while attempting to arrive at some sort of absolute performance standard based upon item content, introduce a great deal of variance into the process through the use of judges. There is a degree of arbitrariness to each of the methods, and thus they should be used cautiously. A suggestion for the use of these methods is to choose one of them and then use it consistently. One should not set a cutting score on one test in an instructional sequence using the Nedelsky method and then use the Ebel method on another test given later. It would seem important that if one is going to set performance standards on the basis of item content that he/she do so consistently. Further, once a method has been chosen (such as the Nedelsky), then the parameters involved in the method should be kept similar over tests in an instructional sequence. For instance, the value of  $K$  in the Nedelsky method should not be varied greatly over testing occasions, or if the Ebel method is used, the levels of the dimensions should be kept the same.

#### 4.2.2 Judgmental Models—Guessing and Item Sampling

In this section, some concerns initially expressed by Millman (1973) concerning errors due to guessing and item sampling will be discussed.

If the test items allow a student to answer questions correctly by guessing, a systematic error is introduced into estimates of student proficiency. There are three ways to rectify this situation:

1. The cut-off can be raised to take into account the contribution expected from the guessing process.
2. A student's score can be corrected for guessing and then the adjusted score compared to the performance standard.
3. The test itself can be constructed to minimize the guessing process.

Methods one and two assume that guessing is of a pure, random nature, which is not likely to be the case for criterion-referenced tests. Thus, adjusting either the cutting score or the student's score will probably prove to be inadequate. The test must be structured to keep guessing to a minimum, because if it occurs, it can't be adequately corrected for.

Also, if because of problems of test construction, inconvenience of administration, or a host of other problems, the test is not representative of the content of the domain, then Millman (1973) suggests that the cutting score be raised (or lowered) an amount to protect against misclassification of students; i.e., false-positive and false-negative errors. Millman offers no methods for determining

the extent or direction of correction for these problems. It is this author's opinion that the test practitioner should exert extra effort to assure that the problem just discussed doesn't occur in the first place. Once again, there doesn't appear to be an adequate method for "correcting away" the problem.

#### 4.2.3 Empirical Models—Data From Two Groups

Berk (1976) has presented a model for setting cutting scores that is based on empirical data. In his paper, he selects empirically the optimal cutting score for a test based upon test data from two samples of students, one of which has been instructed on the material, and the other uninstructed.

Berk offers three ways of approaching the issue of cutting scores based upon the empirical data he collects: (1) Classification of outcome probabilities, (2) computation of a validity coefficient, and (3) utility analysis. He offers a fourth procedure involving incremental validity, but this author feels that procedure has less relevance for the typical situations encountered by the individual setting a standard. In discussing Berk's three methods, we (1) will describe the basic situation common to all three procedures, (2) will offer a series of steps for each of the procedures, and (3) because of the relevance of the procedures, will further discuss one method and offer a graphical solution to the cut-score problem.

### i. The Basic Situation

Two criterion groups are selected for use in this procedure, one group comprised of instructed students and another of uninstructed students. The instructed group should, according to Berk, "consist of those students who have received 'effective' instruction on the objective to be assessed." Berk suggests that these groups should be approximately equal in size and greater than 100 for stable estimates of probabilities. Test items measuring one objective are then administered to both groups and the distribution of scores (putting both groups together) can be divided by a cutting score into two mutually exclusive categories: Predicted mastery and predicted non-mastery.

Combining the classifications of students by predictor (test score) and criterion (instructed vs. non-instructed status) results in four categories that we can represent in a 2 x 2 table, with relevant marginals:

1. True Master (TM): an instructed student whose test score is above the cutting score (C).
2. False Master (FM): a misclassification error where an uninstructed student's test score lies above the cutting score (C).
3. True Non-Master (TN): an uninstructed student whose test score lies below the cutting point (C).
4. False Non-Master (FN): a misclassification error where an instructed student's test score lies below C.

Tabularly, this can be presented as follows. Note how the marginals are defined because they are used in the formulations to follow.

		CRITERION	
		Instructed (I)	Uninstructed (U)
Predictor (Cutting Score)	Predicted Masters PM=TM+FM	(TM)	False-positive Errors (FM)
	Predicted Non-Masters PN=FN+TN	False-negative Errors (FN)	(TN)
		Masters M=TM+FN	Non-Masters N=FM+TN

#### ii. Classification of Outcome Probabilities

In this procedure, identification of the optimal cutting score involves an analysis of the two-way classification of outcome probabilities shown above. This can be done algebraically by following the steps listed below, or graphically, as illustrated in a subsequent section. The steps to follow are:

1. Set up a two-way classification of the frequency distribution for *each* possible cutting score.
2. Compute the probabilities of the four outcomes (for each cutting score) by expressing the cell frequencies as proportions of the total sample. For instance:

$$\begin{aligned}\text{Prob (TM)} &= \text{TM}/(\text{M}+\text{N}) \\ \text{Prob (FM)} &= \text{FM}/(\text{M}+\text{N}) \\ \text{Prob (TN)} &= \text{TN}/(\text{M}+\text{N}) \\ \text{Prob (FN)} &= \text{FN}/(\text{M}+\text{N})\end{aligned}$$

3. For each cutting score, add the probability of correct decisions: Prob (TM) + Prob (TN), and the probability of incorrect decisions: Prob (FN) + Prob (FM).
4. The optimal cutting score is the score that maximizes Prob (TM) + Prob (TN) and minimizes Prob (FN) + Prob (FM). It is sufficient to observe the score that maximizes

Prob (TM) + Prob (TN) because  $[\text{Prob (FN)} + \text{Prob (FM)}] = 1 - [\text{Prob (TM)} + \text{Prob (TN)}]$ . That is, the score that maximizes the probability of correct decisions automatically minimizes probability of incorrect decisions.

### iii. Graphical Solution

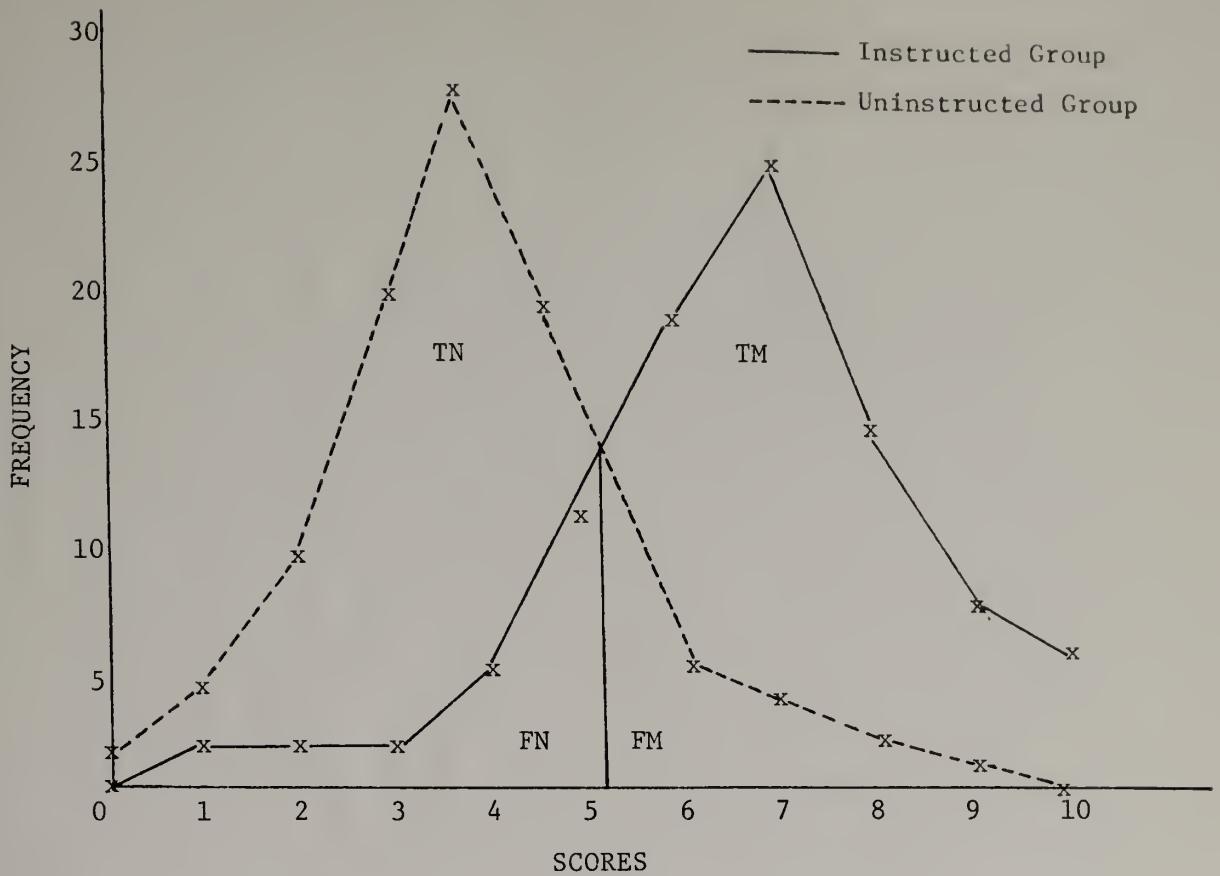
Berk (1976) also mentions that the optimal cutting point for a criterion-referenced test can be located by observing the frequency distributions for the instructed and uninstructed groups.

According to Berk:

The instructed and uninstructed group score distributions are the primary determinants of the extent to which a test can accurately classify students as true masters and true non-masters of an objective. The degree of accuracy is, for the most part, a function of the amount of overlap between the distributions. [p. 5]

If the test score distributions completely overlap, no decisions can be made. The ideal situation would be one in which the two distributions have no overlap at all. A typical situation we should hope for is for the instructed group distribution to have a negative skew, the uninstructed group to have a positive skew, and for there to be a moderate overlap. The point at which the distributions intersect is then the optimal cutting score (C).

For example, suppose we had two groups of 100 students who took a 10 item criterion-referenced test. One group had received instruction, the other had not. A typical plot of the distribution might be:



Score	Uninstr. Freq.	Instructed Freq.
0	1	0
1	5	2
2	10	2
3	20	2
4	30	6
5	20	12
6	6	18
7	4	25
8	3	16
9	1	9
10	0	8

TN: True Non-masters  
 FN: False Non-masters  
 TM: True Masters  
 FM: False Masters

The optimal cutting score is a little greater than 5. Rounded to the integer, 5 would be the optimal cutting score.

#### iv. Validity Coefficient

In this procedure, a validity coefficient is computed for each possible cutting score. The cutting score yielding the highest validity coefficient also yields the highest probability of correct decisions. To utilize the procedure, the following steps should be followed:

1. From the two-way classification (used in procedure 3b) compute the base rate (BR) and the selection ratio (SR). They are given by:

$$BR = \text{Prob (FN)} + \text{Prob (TM)}$$

$$SR = \text{Prob (TM)} + \text{Prob (FM)}$$

2. Calculate the phi coefficient  $\phi_{VC}$  using the following formula:

$$\phi_{VC} = \frac{\text{Prob (TM)} - BR (SR)}{\sqrt{BR (1-BR) SR (1-SR)}}$$

3. The cutting score yielding the highest  $\phi_{VC}$  is the optimal cutting score.

The formula for the phi coefficient,  $\phi_{VC}$ , given above is suitable for a 2 x 2 table of cell probabilities. More generally, the phi coefficient is the Pearson product moment correlation between two dichotomous variables, and could be arrived at as follows:

1. Each student with a test score above the cutting score in question is assigned a 1, below a 0.
2. Each student in the instructed group is assigned a 1, in the uninstructed group, a 0.
3.  $\phi_{VC}$  would then be the correlation coefficient computed in the usual way.



### v. Utility Analysis

In this section, costs or losses are assigned to the misclassification of students as false masters or false non-masters.

Berk notes the following fact:

When the outcome probabilities or validity coefficient approach is used to select the optimal cutting score, it is assumed that the two types of errors are equally serious. If, however, this assumption is not realistic in terms of the losses which may result from a particular decision, the error probabilities need to be weighted to reflect the magnitude of the losses associated with the decision.  
[p. 7]

Berk notes that determination of the relative size of each loss is judgmental, and must be guided by the consequences of the decision considered. He mentions considering the following factors: student motivation, teacher time, availability of instructional materials, content, and others. Berk suggests the following, which we have capsulized into a series of steps:

1. Estimate the *expected disutility* of a decision strategy ( $\zeta$ ) by

$$\zeta_k = \text{Prob (FN)}[D_1] + \text{Prob (FM)}[D_2]$$

where  $D_1$  and  $D_2 < 0$

and  $k$  = the single decision in question

$D_1$  and  $D_2$  = respective disutility values

2. Estimate the *expected utility* of a decision strategy ( $v$ ) by

$$v_k = \text{Prob (TM)}[U_1] + \text{Prob (TN)}[U_2]$$

where  $U_1$  and  $U_2 > 0$

and  $k$  = the single decision in question (same as for disutility)

$U_1$  and  $U_2$  = respective utility values

3. Form a composite measure of test usefulness by combining the estimates of utility and disutility across all decisions.

$$\gamma = \sum_{k=1}^n (v_k + \zeta_k)$$

$\gamma$  = index of expected maximal utility.

4. Choose the cutting score with the highest  $\gamma$  index (it maximizes the usefulness of the test for decisions with a specific set of utilities and disutilities).

#### 4.2.4 Decision-Theoretic Procedures

Berk (1976) has looked at the minimization of false-positive and false-negative decisions through the use of actual test data. He selects as optimal the cutting score that minimizes false-positive and false-negative errors. Another way to look at false-positive and false-negative errors is to assume an underlying distributional form for your data and then observe the consequences of setting values, such as cutting points, based upon the distributional model. The logic is the same here in terms of minimization of errors, except that by assuming a distributional form, actual data does not have to be collected. Situations can be simulated or developed, based upon the model.

Meskauskas (1973) has related and compared these procedures to those based upon analyses of the content of the test. In reference to these models, of which we will describe one:

. . .the models to follow deal with approaches that start by assuming a standard of performance and then evaluating the classification errors resulting from its use. If the error rate is inappropriate, the decision-maker adjusts the standard a bit and tries his equation again. [p. 139]

Before discussing one of the procedures in greater detail, the Kriewall binomial-based model, the procedures discussed here should be related to criterion-referenced testing procedures involving the determination of test length. Many of the test length determination procedures (Millman, 1973; Novick & Lewis, 1974) make underlying distributional assumptions and proceed in the fashion discussed above by Meskauskas. The focus of concern, however, is test length determination, and not the setting of a cutting score. In fact, Millman's (1973) procedure is based upon exactly the same underlying distribution, the binomial, as is Kriewall's model to be discussed. All that differs is the focus of concern. It should be pointed out that the procedures are exactly the same, the data is just represented differently because of the level of concern, either cutting score or test length.

#### i. Kriewall's Model

Kriewall's (1972) model focuses on categorization of learners into several categories: non-master, master, and an in-between state where the student has developed some skills, but not enough to be considered a master. We see here the first mention of an "indifference zone," critical to the work of Phaner (1974) and Wilcox (1976) on test length. Thus, Kriewall assumes the function of measurement, using the test, is to classify students into one of two categories, master or non-master. Of course, the test, as a sample of the domain of tasks, is going to misclassify some individuals as false-positives and false-negatives. By assuming a particular distribution, these errors may be studied.

Kriewall's probability model, used to develop the likelihood of classification errors, is based upon the binomial distribution.

He assumes:

1. The test represents a randomly selected set of dichotomously scored (0-1) items from the domain.
2. The likelihood of correct response for a given individual is a fixed quantity for all items measuring a given objective.
3. Responses to questions by an individual are independent. That is, the outcome of one question is independent of the outcome of any other question.
4. Any distribution of difficulty of questions (for an individual) within a test is assumed to be a function of randomly occurring erroneous responses (Meskauskas, 1976).

With these assumptions, Kriewall views a student's test performance as "a sequence of independent Bernoulli trials, each having the same probability of success." A sequence of Bernoulli trials follows a binomial distribution, which has a probability function which relates the probability of occurrence of an event (a particular test score) to the number of questions in the test by:

$$f(x) = \binom{n}{x} p^x q^{n-x} ,$$

where

$x$  = a test score

$n$  = total number of test items

$p$  = examinee domain score

$q$  =  $1-p$

and

$$\binom{n}{x} = \frac{n!}{x! (n-x)!} .$$

Kriewall sets some boundary values and a cutting score, and then looks at the probability of misclassification errors. Using the notation of Meskauskas (1976), set:

$z_1$  = the lower bound of the mastery range (as a proportion of errors)

$z_2$  = the upper bound of the non-mastery range

$C$  = the cutting score; the maximal number of allowable errors for masters. Kriewall recommends

$$C = \frac{z_1 + z_2}{2} .$$

Given values for the above three variables, Kriewall uses the (assumed) binomial distribution to determine the probabilities. If  $\alpha$  is the probability of a false-positive result and  $\beta$  is the probability of a false-negative result, then  $\alpha$  and  $\beta$  are given by:

$$\alpha = \sum_{w=c}^n \binom{n}{w} z^{n-w} (1 - z_1)^w$$

$$\beta = \sum_{w=0}^{c-1} \binom{n}{w} z^{n-w} (1 - z_2)^w$$

where  $w$  = observed number of errors (and  $w = n-x$ ) for an individual.

According to Meskauskas (1976) the formula for  $\alpha$  is:

. . . equivalent to obtaining the probability that, given a large number of equivalent trials, a person whose true score is equal to the lowest score in the mastery range will fall in the non-mastery range. [p. 141]

By setting  $z_1$  and  $z_2$  at various values, and determining  $C = \frac{z_1 + z_2}{2}$ , the probabilities of false-positive and false-negative errors can be studied. The optimal value for  $C$  (and thus  $z_1$  and  $z_2$ )

would then be the value that minimized  $\alpha$  and  $\beta$ . The results are dependent, however, on  $n$  and  $w$ .

### ii. Suggestions

While Kriewall has offered us a method of studying classification errors that does not depend upon actual data, this author prefers the method of Berk, due to its simplicity. Kriewall's model seems to this author to fit in much better with the procedures on test length determination. For instance, suppose you have specified minimal values for  $\alpha$  and  $\beta$ , and have determined  $C$ , the cutting point. Then the formulas above for  $\alpha$  and  $\beta$  can be solved for  $n$ , the total number of questions needed. (It would be much easier if one isolated  $n$  on the left hand side.) This is exactly what is done when using the binomial model to solve the test length problem.

In sum, we prefer the Berk method for observing probabilities of misclassification errors both because of its simplicity and because of the lack of restricting underlying distributional assumptions. Kriewall's method does, however, offer a viable alternative for setting a cut-off score when actual test data cannot be collected.

#### 4.2.5 Empirical Models Depending Upon a Criterion Measure

The models to be discussed in this section bear great resemblance to both Berk's and Kriewall's methods just discussed. They have been separated from those two methods because these methods are built upon the existence of an outside criterion measure,

performance measure, or true ability distribution. The test itself, and the possible cut-off scores, are observed in relation to this outside measure. An optimal cut-off is then chosen in reference to the criterion measure. For instance, Livingston's (1975) utility-based approach leads to the selection of a cut-off score that optimizes a particular utility function. The procedure of Vander Linden and Millenburgh (1976), in contrast, leads to the selection of a cut-off score that minimizes expected loss.

In sum, to utilize these procedures, a suitable outside criterion measure must exist. Success and failure (or probability of success and failure) is then defined on the criterion variable and the cut-off chosen as the score on the test that maximizes (or minimizes) some function of the criterion variable. The existence of such a criterion variable has implications for that utilization of these methods for setting cut-off scores on minimum competency tests. This will be discussed in greater detail later in this chapter.

#### i. Livingston's Utility-based Approach

Livingston (1975) suggests the use of a set of linear or semi-linear utility functions in viewing the effects of decision-making accuracy based upon a particular cut-off score. That is, the functions relating benefit (and cost) of a decision are related linearly to the cutting score in question.

Livingston's procedure is like Berk's procedure for utility analysis discussed in 4.2.3, except that Livingston develops his

procedure based upon any suitable general criterion measure (not just instructed versus uninstructed), and also specifies the relationship between utility (benefit or loss) and cutting scores as linear. The relationship does not have to be linear; however, using such a relationship simplifies matters somewhat. In such a situation the cost (of a bad decision) is proportional to the size of the errors made and the benefit (of a good decision) is proportional to the size of the errors avoided.

Rather than discuss this procedure further at this time, it can be recommended that when a decision has been made to utilize a utility-based approach and the outside criterion is not clearly specified as instructed versus uninstructed, that Livingston's method be consulted. We should note that in his paper, he develops the procedures for non-linear functions first, and then considers linear functions as a special case.

ii. Van der Linden and Mellenburgh's  
Approach

The developers of this procedure have prescribed a method for setting cutting scores that is related both to Berk's procedure and Livingston's. We will describe the procedure briefly and in the process relate it to Berk's work. A test score is used to classify examinees into two categories: accepted (scores above the cutting score) and rejected (scores below). Also, a latent ability variable is specified in advance and used to dichotomize the student population: students above a particular point on the latent variable are considered "suitable" and below "not suitable." The situation may be represented as follows:



		Latent Variable	
		Not suitable $\gamma < d$	Suitable $\gamma \geq d$
Decision	Accepted $X \geq C$	"False +" $l_{01}(\gamma)$	$l_{11}(\gamma)$
	Rejected $X < C$	$l_{00}(\gamma)$	"False -" $l_{10}(\gamma)$

where  $C$  = cutting score on the test,

$d$  = cutting score on the latent variable ( $0 \leq d \leq 1$ ),

and where  $l_{ij}$  ( $k, j = 0, 1$ ) is a function of  $\gamma$  and related in the general loss function:

$$L = \begin{cases} l_{00}(\gamma) & \text{for } \gamma < d, X < C \\ l_{10}(\gamma) & \text{for } \gamma \geq d, X < C \\ l_{01}(\gamma) & \text{for } \gamma < d, X \geq C \\ l_{11}(\gamma) & \text{for } \gamma \geq d, X \geq C \end{cases}$$

The authors then specify risk (the quantity to be minimized) as the expected loss, and the cutting score that is optimal is the value of  $C$  that minimizes the risk function (expected value of loss). They simplify matters (as does Livingston) by specifying their loss function as linear.

In sum, while van der Linden and Mellenburgh have provided a method for setting a cut-off score on the test, they have offered little to help in setting the cut-off on the latent variable. In a sense then, they have only transferred the problem of a cut-off to a different measure, and hence "begged" the question at hand. It should be noted that this procedure accentuates the problem more

than the other procedures being discussed here. All these methods switch the cut-off problem from the test itself to an outside criterion measure.

iii. Livingston's Use of Stochastic  
Approximation Techniques

Livingston (1976) has developed procedures for setting cut-off scores based upon stochastic approximation procedures. According to Livingston, the problem involving cut-off scores can be phrased as follows to fit stochastic procedures: "In general, the problem is to determine what level of input (written test score) is necessary to produce a given response (performance), when measurements of the response are difficult or expensive." The procedure, according to Livingston, is as follows:

1. Select a person; record his/her test score and measure his/her performance.
2. If the person succeeds on the performance measure (if his/her performance is above the minimum acceptable), choose next a person with a somewhat lower test score. If the person fails on the performance measure, choose a person with a higher written test score.
3. Repeat step 2, choosing the third person on the basis of the second person's measured performance.

Livingston offers two different procedures for choosing step size, the up-and-down and the Robbins-Monro Procedure, and a number of procedures for estimating minimum passing scores consonant with each.

This procedure, like those discussed earlier in this section, depends upon the existence of a cut-score established on another variable, this time the performance measure, in order to establish

the passing score on the test. This then limits greatly the applicability of the method. Livingston (personal communication, 1978) has suggested that judgmental data on performance can be used, rather than actual performance data, with the procedure, but this has yet to be documented in any fashion. When documented, the possibilities for use of the procedures will be greatly expanded.

#### iv. Huynh's Procedures

Huynh (1976), and more recently, Huynh and Perney (1977), have advanced procedures for setting cut-off scores that are predicated on the existence of a criterion measure or referral task. This referral task can be envisioned as an external criterion to which competency can be related. For instance, Huynh (1976) states that "Mastery in one unit of instruction may not be reasonably declared if it cannot be assumed that the masters would have better chances of success in the next unit of instruction." The next unit in this case would be the criterion measure. Huynh and Perney (1977) have taken the general procedures suggested by Huynh (1976) and applied them to the case when instructional units are sequenced in a linear hierarchy. This simplifies somewhat the mathematical complexity of the formulations.

These procedures once again depend upon an outside criterion variable to allow the estimation of a cut-score on the test. In this case, the user of the method is asked to establish the probability of success of individuals on the referral task. Because of the necessity of a criterion variable for operation, these

procedures suffer in generalizability. They are, for instance, apparently not useful for minimum competency testing situations where a criterion variable, and associated probability of success, are next to impossible to establish.

#### 4.2.6 Educational Consequences

In this situation, one is concerned with looking at the effect setting a standard of proficiency has on future learning or other related cognitive or affective success criteria. According to Millman (1973), the question here is "What passing score maximizes educational benefits?".

This approach can be visualized from an experimental design point of view. A subject matter domain is taught to a class of students who are then tested on the material. These students are assigned (randomly) to groups with the groups differing on the performance level required for passing the test. The students are then assessed on some valued outcome measure and the level of performance on the criterion-referenced test for which the valued outcome is maximal (it could be a combination of valued outcomes) becomes the performance standard or criterion score.

Thus, to use this method, much more data needs to be collected than for the item content procedure. You really have to run an experiment, and then your performance standard based upon the results of the experiment. It is for this reason that we feel that the procedure offers less potential use for the practitioner concerned about setting a performance standard or cutting score. One seldom

has control over variables necessary for setting up experiments based upon a "true research" paradigm.

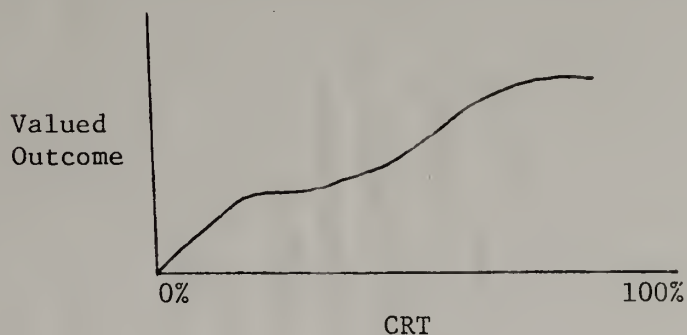
### i. Block's Study

Block's study (1972) involves students learning a subject segment on matrix algebra using a Mastery Learning paradigm. Such a paradigm dictates that students who don't perform adequately on the posttest be recycled through remedial activities until they demonstrate mastery (i.e., attain a score above the cutting score). Block established four groups of students, where each group was tested using one of the following four performance standards: 65, 75, 85, and 95% of the material in a unit must be mastered before proceeding on the next unit. He then examined the effects of varying the performance standard on six criteria that were used as the variables to be maximized. Viewing these criteria as either cognitive or affective, Block observed that the 95% performance level maximized student performance on the cognitive criteria, while the 85% performance level seemed to maximize the affective criteria.

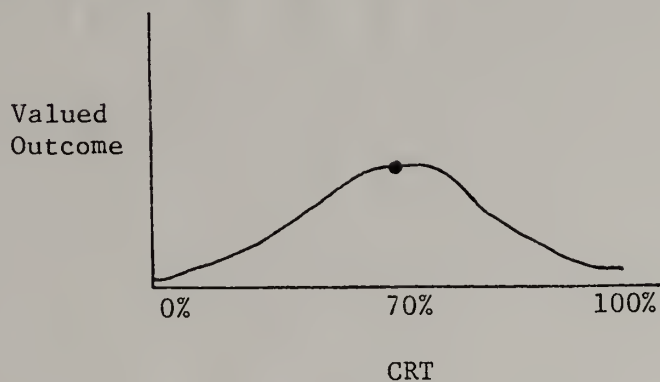
Some comments on Block's study are in line. One, the results lack generalizability. The 95% and 85% levels, which maximize the cognitive and affective measures respectively, are likely to change with the subject matter. Two, as pointed out by Glass (1978a), the method of maximizing a valued outcome assumes that there is a distinct point or criterion score on the CRT that maximizes the outcome. What if the curve relating performance on the CRT is monotonically increasing, so that 100% performance on the CRT maximizes the valued

outcome? In fact, this author agrees that it is more likely to be the case that the graph is nonotonically increasing than the case where the graph increases and decreases. For example (Glass, 1978a):

1. Monotonically increasing graph (Problem situation)



2. Ideal situation



Thus, it can be seen that unless the graph increases and then decreases, a 100% performance standard will be optimal. This standard is of limited use because it is not realistic to expect all students to attain that level.

Third, Block discusses that if there are multiple criteria to be maximized as valued outcomes, then some model for combining criteria with relevant weights needs to be developed. He does not

offer any procedures for doing so however, and he looks at the effects of the performance standards on each of the 6 criteria separately. It should be noted that multiple criteria is a way around the problem discussed above (Glass, 1978a). For instance, if one of the valued outcomes has a monotonically increasing relationship with the test scores and the other monotonically decreasing relationship, then the composite should have a peak value at a point other than 0% or 100%. While this would seem to solve the problem, another problem is only further exacerbated; what weights should be assigned to the valued outcomes to form the composite? These procedures have not yet been developed, and further, they are likely to be situation specific.

#### 4.2.7 Combination Models: Judgmental-Empirical

Zieky and Livingston (1977), and more recently, Popham (1978b), have suggested two procedures that are based upon a combination of judgmental and empirical data. In addition, Zieky and Livingston have included an in-depth discussion of how to implement the procedures, something that has been lacking with many other procedures. The two procedures presented by Zieky and Livingston, the Borderline-Group and Contrasting-Groups methods, are procedurally similar. They differ in the sample of students on which performance data is collected. Further, while judgments are required, the judgments necessary are on students; not on items, as are many of the other judgmental methods (Nedelsky, Angoff, Ebel, etc.). Zieky and

Livingston make the case that judging individuals is likely to be a more familiar task than judging items. Teachers are the logical choice as judges, and for them, the assessment of individuals is commonplace.

#### i. Borderline-Group Method

This method requires that judges define what they would envision as minimally acceptable performance on the content area being assessed. The judges are then asked to submit a list of students (about 100 students) whose performances are so close to the borderline between acceptable and unacceptable that they can't be classified into either group. The test is thus administered to this group, and the median test score for the group is taken as the standard.

#### ii. Contrasting-Group Method

Once judges have defined minimally acceptable performance for the subject area being assessed, the judges are asked to identify those students they are sure are either definite masters or non-masters of the skills measured by the test. Zieky and Livingston suggest one hundred students in the smaller group in order to assure stable results. The test score distributions for the two groups are then plotted and the intersection is taken as the initial standard. This is exactly the same as the graphical procedure suggested by Berk, and presented in section 4.2.3. Zieky and Livingston then suggest adjusting the standard up or down to reduce false masters or false non-masters.



### iii. Suggestions

These methods, particularly the Contrasting-Groups Method, are very similar to the procedure suggested by Berk. Instead of actually forming instructed and uninstructed groups, however, as suggested by Berk, the Contrasting-Groups Method asks judges to form the groups. This judgmental procedure would seem more advantageous when the content being assessed has had a long instructional period (minimum competency testing is an example), or when there would be problems justifying the existence of an uninstructed group. Berk's method would be more useful for tests based on short instructional segments, most likely administered at the classroom level.

A comparison of the judgments involved in the two procedures indicates that the Contrasting-Groups Method would be the most easy method to justify using. It is a more reasonable task to identify "sure" masters and non-masters than it is to identify borderline students in the subject area being assessed. In sum, the Contrasting-Groups Method appears to this author to be a most reasonable way of setting a performance standard.

#### 4.2.8 Combination Models: Bayesian Procedures

Novick and Lewis (1974) were perhaps the first to suggest that Bayesian procedures are useful for setting standards. Schoon, Gullion, and Ferrara (1978) have recently reviewed Bayesian procedures. According to Hambleton and Novick (1973), Bayesian procedures allow the incorporation of:

1. A loss ratio, reflecting the severity of false-positive and false-negative errors,
2. prior information on examinee domain scores,
3. test score information, and
4. a decision criterion; the degree of certainty that an examinee's domain score exceeds a cut-off score.

A cut-off score must first be set in order for the four factors to be incorporated. Thus, Bayesian procedures offer a way of augmenting the establishment of a cut-off score rather than a method for setting the cut-off score itself.

#### 4.3 Setting Cut-Scores for Minimum Competency Tests<sup>1</sup>

The minimum competency movement, replete with all its problems, is now a reality. While much could be said about the philosophical, psychological and legal implications of this movement, for the purposes of this dissertation only the problem of setting cut-off scores on minimum competency tests will be addressed.

Hambleton and Eignor (1979) offer the following definitions of a minimum competency test:

A minimum competency test is designed to determine whether an examinee has reached a prespecified level of performance relative to each competency being measured. The "prespecified level" or "standard" may vary from one competency to the next. Also, each competency is described by a well-defined behavior domain.

---

<sup>1</sup>Some of the material in this section is from a paper by Hambleton and Eignor (1978c).

A "standard" is a point on a test score scale which is used to separate examinees into two categories, each reflecting a different level of proficiency relative to the competency measured by the test under consideration. It is common to assign labels such as "master" or "competent" to those persons in the higher-scoring category and "non-master" or "incompetent" to those persons in the lower-scoring category. Note that if a test measures more than a single competency and if examinees are to be classified into competency categories based on their performance on each set of items measuring a competency, as is often the case, a standard is set for each competency measured by the test. There will be an many competency decisions as there are competencies measured by the test.

Given the definition of minimum competency testing just presented, an important question becomes which of the cut-score methods presented in Figure 4.2.2 are applicable? As the ensuing discussion will indicate, many of these methods, for a number of reasons, simply are not applicable. The first area to be discussed concerns the empirical models, and as will be shown, none of these are really suitable for minimum competency standard setting.

The methods for setting cut-off scores that depend upon the existence of a criterion measure, performance measure, or latent ability continuum (Livingston, 1975; Livingston, 1976; Huynh, 1976; Huynh and Perney, 1977; Van der Linden & Mellenburgh, 1977) are difficult to apply in this situation. This is because any external criterion variables that would be appropriate for validating minimum competency tests are going to be difficult to gain agreement

about and very difficult to measure. Take the case of high school minimum competency testing as an example. In this case the criterion variable is most often discussed as "life success." How does one go about operationally defining "life success" and then measuring it? Reading experts, for instance, are not going to have the same idea about what the minimally competent person can read. Should he/she be able to read on the 4th, 8th, or 12th grade level? More concisely, using Jaeger's (1978) example:

"Educators would no sooner agree on the proportion of New York Times front page passages eleventh-graders should be able to comprehend and explain than they would the proportion of multiple-choice test items those eleventh-graders should answer correctly, so as to be labeled 'minimally competent'." Thus, the jist of this reasoning is that if agreement cannot be reached on the criterion measure, then methods for setting standards that depend upon a criterion measure are not workable.

The decision-theoretic procedure offered by Kriewall (1972) is also difficult to apply in this context. This procedure is based upon the definition of (usually) two mastery states. The cut-off on the test is then selected as the point that minimizes false-positive and false-negative errors in the classifying of individuals into the defined mastery states. Once again, the problem is evident. The mastery categories would in this case be "competent" and "incompetent," and they are essentially undefined. Until people can agree on a definition of minimum competence, it is not possible to use this or other decision-theoretic procedures.

You cannot minimize errors of prediction if the categories to be predicted can't be established. Jaeger (1978) makes the same point, and then goes on to say that while the models allow for different utilities to be associated with false-positive and false-negative errors, no guidelines exist for establishing these values. Thus, if the categories to be predicted could be established, other problems would have to be contended with.

Berk's method (1976), based on an instructed and an uninstructed group, has a problem when applied conceptually to minimum competency testing. There is simply no reasonable, or ethical, way of establishing groups instructed on minimum competency skills from groups that have been withheld instruction. Because minimum competency testing involves skills that are developed over a period of years, it is simply impossible to justify withholding the instruction for the group by claiming they will be instructed after the standard is set. Berk also suggests using a pretest and posttest procedure on the same group to form the uninstructed and instructed groups. The problem with this approach is the time interval; changes in performance could be attributed to any number of other variables besides instruction.

Block's method (1972) depends upon the maximization of some valued outcome measure. Again, when applied to minimum competency testing, the problem is evident. One can't maximize a valued outcome if the outcome can't be defined in any reasonable manner in the first place. To utilize Block's method, there would have to be concensual agreement on what a valued outcome of being competent

is. This would seem to be as difficult a task as trying to get people to define behavior associated with minimum competency.

Other suggestions for setting cut-off scores come under scrutiny when applied to minimum competency tests. For instance, Millman's (1973) suggestion about adjusting the cut-off for the effects of guessing is an example. Educational Testing Service has corrected the cut-points on the National Teachers Exams (1976) to take care of guessing. The problem here is that for minimum competency tests, pure, random guessing rarely occurs, and because of this, the effects of raising the cut-off scores as if it had are unknown. Research in this area is badly needed.

Figure 4.3.1 represents that portion of Figure 4.2.2 that is applicable to minimum competency tests. (Bayesian procedures have also been deleted from Figure 4.3.1 because a cut-off score must first exist to utilize these procedures.) Table 4.3.1 provides a comparison of the methods that are suitable, in this author's judgment, for setting standards on minimum competency tests. What is not explained in Table 4.3.1 is why the Modified Angoff and Contrasting Groups procedures were selected by this author as the two most suitable methods to use for setting minimum-competency standards. What follows is a brief critique of the other methods. Hopefully this will serve as an initial justification for the choices made.

Recent interactions have lead this author to question the Nedelsky method on two levels. One, discussions with practitioners in the field have pointed out that the Nedelsky method is difficult

Judgmental Models

<u>Item Content</u>	<u>Judgmental-Empirical</u>
Nedelsky (1954)	Contrasting-Groups
Modified Nedelsky	Borderline-Group
Angoff (1971)	
Modified Angoff	
Ebel (1972)	
Jaeger (1978)	

Figure 4.3.1 A classification of models and methods for determining minimum competency standards.

Table 4.3.1

## A Comparison of Several Standard Setting Methods

Question	Judgmental						Combination	
	Modified Nedelsky	Angoff	Modified Angoff	Ebel	Jaeger	Contrasting Groups	Borderline Group	
1. Is a definition of the minimally competent individual necessary?	Yes	Yes	Yes	Yes	No	No	Yes	
2. What is the nature of the rating task— or items, or individuals?	Items	Items	Items	Items	Items	Individuals	Individuals	
3. Are examinee data needed?	No	No	No	No	No	Yes	Yes	
4. Do judges have access to the items?	Yes	Yes	Yes	Yes	Yes	Usually, but don't need to	Usually	
5. Are the judgments made in a group setting or individual setting?	Both	Both	Both	Both	Both	Individual	Individual	
Choices of methods to use for setting standards on minimum competency tests.			✓			✓		



to implement. Further, these discussions have demonstrated that at times, the method for arriving at the cut-point was done improperly. Secondly, the method does not allow a minimum passing level for an item to be in the interval from .51 to .99. For instance, for a five choice item, the possible points a judge can choose are:

0        .25        .33        .50        1.00

Other methods, such as Angoff's, allow the choice points to vary on the whole interval from 0 to 1.0. While research in this area is needed, it would seem to this author that a model that allows use of the whole continuum, or points interspersed on this continuum, would be preferred.

The Modified Nedelsky procedure, while probably applicable for certain situations, does not provide the detailed data necessary for setting standards on minimum competency tests. Given the fragile legal status of such testing, a usable method that incorporates more choices for the judges, and thus more data, would seem preferable.

Educational Testing Service has developed the Modified Angoff procedure from a concern that Angoff's initially suggested procedure may be overly difficult for judges. Given that the judge is, at best, providing a "ball-park" estimate, why not provide him/her with some fixed scale points to operate with, and thereby simplify the task?

Much has been said of a cautioning note, both in this chapter and elsewhere, about Ebel's procedure. Worthy of reiteration at this point is the arbitrary nature of the grid and the lack of accompanying guidelines for choice of type and number of dimensions. Ebel's method, without further guidelines, is very difficult for the practitioner involved with minimum competency to implement.

Jaeger's method, to the best of this author's knowledge, has yet to be field-tested in any formal fashion. While conceptually it appears to offer a viable approach, it is yet to be in a form that a practitioner could implement.

Finally, the Borderline-Group method was eliminated from the list because of the type of judgment the judges are asked to make. It seems to this author much more reasonable to ask judges to select definite masters and non-masters of the content being assessed (Contrasting-Groups Method) than it is to ask them to select a group of borderline students.

In sum, if placed in a position of having to assist in setting a standard for a minimum competency test, this author would utilize either the Modified Angoff or Contrasting-Group Method. The final choice would depend upon whether or not the judges personally knew a group of individuals taking the test. The selection of these two methods is based upon a rational analysis of the available methods.

#### 4.4 An Application of the Modified Angoff Procedure

One of the tasks assigned to this author upon commencement of employment at Educational Testing Service in September 1978 was to determine and implement a suitable procedure for setting a cut score on the Multistate Insurance Licensing Program (MILP) certification examinations. These exams are four in number, and they qualify as minimum competency exams. Some background information will be provided next about the Program and the tests, followed by a discussion of the cut-score method chosen and the results of the application of the procedure.

In the context of the research presented in this dissertation, the cut-score procedures to be discussed can be viewed as a field application of one of the suggested procedures, the Modified Angoff technique.

##### 4.4.1 Background Information

The Multistate Insurance Licensing Program (MILP) was developed beginning in 1974 under the sponsorship of the National Association of Insurance Commissioners. The licensing tests, developed by ETS, cover the four major lines of insurance: (1) life, (2) accident and health, (3) property, and (4) casualty. Each of the four tests has two major parts:

1. Part 1 consisting of 50 multiple-choice questions covering basic principles of insurance and product knowledge that is uniform across all states.

2. Part II consisting of 25-40 multiple-choice questions covering individual state laws, rules, and regulations plus subject matter unique to the state.

The first examinations were administered in Illinois in October of 1975. Presently seven states participate in MILP: Colorado, Delaware, Illinois, Indiana, Massachusetts, Pennsylvania, and Wisconsin. The tests are administered twice a month with the 1978-79 candidate volume estimated at 45,000. There are a number of forms for the Part I tests, all of which are equated back to a base test, and placed on a common scale. Part II tests are equated, but this is done on a within state basis.

The Part I tests for the four areas, which is the focus of this research, are developed from a set of test specifications. These specifications are detailed content outlines of the topics and subtopics to be covered on the examinations. Task forces made up of insurance commissioners, insurance attorneys, and key industrial experts met in 1974-75 to formulate, revise and finally approve the outlines upon which the tests are based. The charge of their task forces was to develop examinations that would test new agents on the critical subject matter necessary to protect the public welfare. Hence, the specifications, and initial questions, were developed specifically to cover the most basic concepts in insurance and laws, rules, and regulations at the level of minimum competency. The concern was what the beginning minimally competent insurance agent would need to know to "protect the public good." The test specifications are presented in the Appendix for the four tests being discussed.

Periodically, committees of insurance agents, commissioners, and industrial representatives meet to assess the content validity of the items that will constitute a new test form. Part of the research presented in this chapter, though not the focus of the major topic being discussed, concerns the content validation method utilized with the new test forms being considered.

For the seven states that participate in MILP, the Commissioners of Insurance have the statutory responsibility for determining the level of competency that candidates must achieve in order to be granted insurance licenses. Realizing the potential arbitrariness of such a decision, the Commissioners requested assistance in evaluating four new examinations in order to determine the appropriate cut-score for making pass-fail decisions. In addition, it was requested that the new test forms to be used be assessed in reference to content validity.

Given the requests from the Commissioners of Insurance and MILP, two specific goals dictated the procedures that the ETS group working on insurance were to formalize:

1. To select a method and report subsequent information that would assist the seven state commissioners in setting cut-off scores for candidates seeking licenses to sell insurance.
2. To assess the appropriateness of the question themselves as adequate samples of the content domain of the tests, as well as to assess the appropriateness of the content domains for each test as representing knowledge that new agents must possess to insure the public welfare and protect their clients.

The task of deciding upon and implementing a method for setting cut-off scores was initially assigned to this author. The final decision was to be made by the Program Director, Test Development Specialist,

and Statistical Coordinator (this author). The charge of developing methods for assessing the content validity of the test was the task of the Test Development Specialist. However, because the content validity assessment procedures and the cut-score procedures were so closely intertwined, both will be reported upon in this chapter.

Given the concern expressed by the National Association of Insurance Commissioners that a suggested cut-score be offered for each of the validated four Part 1 tests to be administered in November of 1978, a letter was sent out on October 6, 1978 by the Program Director to a selected group of judges representing the seven constituent states. This letter, which gave an overview of the tasks, is containing in the Appendix of this dissertation. Also, at the time, judges were informed of a meeting they were to attend on October 17 at O'Hare Airport in Chicago to set the cut-off scores on the four Part 1 examinations and to assess the content validity of the tests. With the introductory letter, panel members were also sent a set of materials that they would be using at the October 17 meeting. These materials are presented in the Appendix. They included an Overview of Tasks, instructions for completing the Content Rating Form, a sample Content Rating Form, instructions for completing the Question Rating Form, and a sample of a Question Rating Form.

#### 4.4.2 Choice of a Method for Determining a Cut-Score

The choice of the method for setting a cut-score for each of the Part 1 tests was the initial responsibility of this author in the role of Statistical Coordinator. Suggestions were entertained from other knowledgeable staff members of COPA (Center for Occupational and Professional Assessment) and elsewhere at ETS. The initial reaction was to suggest using the Contrasting Groups Procedure discussed earlier in this chapter. It became immediately apparent, however, that the panel of judges would not know a sample of candidates, and thus, a purely judgmental procedure was a necessity. The choice came to that of the Nedelsky procedure or a Modified-Angoff procedure. At first the possibility of utilizing both procedures was considered, but this was dismissed because: (1) there simply would not be enough time to complete a content validation and two cut-off procedures (each judge had to assess 100 items for content validity and cut-off determination), and (2) there existed a deep-seated concern on the part of this author that the two procedures might give very disparate results. The task of "explaining away" differences such as those that surfaced in the Andrew and Hecht (1976) study was to be avoided.

The final choice of a Modified-Angoff approach was made based upon the discussion presented in the previous section of this chapter and also upon a past precedent. ETS had successfully utilized the Modified-Angoff Procedure both in setting cut-scores for the National Teacher Examination (1976) and for setting a cut-score on the MILP exam for the state of Wisconsin (1977). The documentation

for both of these applications was to prove most useful both in terms of suggestions for proceeding and in terms of offering some sort of research foundation.

As a means of double-checking the results of the Modified-Angoff Procedure, another procedure for setting a cut-score was developed to be used in conjunction with the content validation procedures. This procedure was to serve as a check on the Modified-Angoff Procedure and also be offered to the Commissioner as a low priority piece of supplemental information. The tests were subdivided into major content areas and the judges were asked to estimate how many of the questions, representing a content area, would a minimally competent candidate be able to answer correctly? For instance, for the Accident/Health Test, the following four questions were asked:

1. Of the ten questions covering Basics in Accident and Health insurance, how many of these questions do you think a minimally competent person would answer correctly? \_\_\_\_\_
2. Of the twenty questions covering Individual Accident and Health Provisions, how many of these questions do you think a minimally competent person will answer correctly?  
\_\_\_\_\_
3. Of the fifteen questions dealing with Types of Coverage, how many of these questions do you think a minimally competent person will answer correctly? \_\_\_\_\_
4. Of the five questions dealing with Types of Contracts, how many of these questions do you think a minimally competent person will answer correctly? \_\_\_\_\_

Under more general circumstances, such a procedure would be highly questionable at best. The results of making assessments on



such a large number of items should lead to great variation across judges. However, in the present context, where judges have just looked carefully at the content validity of each item, and also looked at the major content sections, such a procedure may not be so unreasonable if used as an independent check. (The results seem to suggest just this point.)

#### 4.4.3 Panel System Design

Because of the nature of the four lines of insurance being considered, life, accident and health, property and casualty, some pairings of the areas seemed in line. For instance, the life and accident and health areas have many common philosophical "roots," sharing basic concepts like parts of a policy, sources of insurability information, representation, and warranties, etc. For many similar reasons, it was decided that the property and casualty areas formed a natural pairing. Thus, it was decided that one panel should be designated to evaluate the life and accident and health tests and a second panel to evaluate the property and casualty tests. The life-accident/health panel was asked to come to a morning meeting on the 17th of October, the property-casualty panel to an afternoon meeting on that date.

The two panels (AM: life-accident/health; PM: property-casualty) were asked to make both a content validity assessment and a cut-score assessment, both to be described later in this chapter. The work for each test was divided into two separate questionnaires and a decision was made to split each panel in half. One half-panel

would assess content first and then answer the question rating form for setting a cut-off; the other half-panel vice-versa.

In this way, a counterbalancing effect would occur in the order of judgments requested. Figure 4.4.1 represents a pictorial of the panel set-up.

In determining the number of individuals to serve on a panel, the following considerations were relevant. One, the diversity of occupational and educational settings within the profession needed to be represented. Two, the need for a sufficient number of judges to obtain reliable judgments and to form half-panels needed to be considered. Third, the probable availability of persons in the insurance areas had to be considered. Given these considerations, it was decided that panels of 16 members each be formed for the life-accident/health and property-casualty areas. Half-panels would then be eight members each. However, because of the availability and willingness of certain individuals qualified in all four areas to participate, the panels sizes were larger. (Seventeen people assessed life, fifteen accident/health, and eighteen each for property and casualty. Only fifteen assessed accident/health because two individuals had to be excused to make plane connections.)

#### 4.4.4 Panel Tasks—Question Rating Form

Prior to the question rating task, four activities took place. First, the panel members were presented with a short twenty minute discussion of cut-scores and how they are used. This author gave the presentation, and was careful to present the need for cut-scores

Life-Accident/Health (AN)		Property-Casualty (PN)	
Half-Panel 1	Half-Panel 2	Half-Panel 1	Half-Panel 2
1. Content Form-Life	1. Question Form-Life	1. Content Form-Property	1. Question Form-Property
2. Question Form-Life	2. Content Form-Life	2. Question Form-Property	2. Content Form-Property
3. Content Form-Accident/Health	3. Question Form-Accident/Health	3. Content Form-Casualty	3. Question Form-Casualty
4. Question Form-Accident/Health	4. Content Form-Accident/Health	4. Question Form-Casualty	4. Content Form-Casualty

Figure 4.4.1 Pictorial representation of panel formulation and sequential order of tasks.

and a brief rationale about why the Modified-Angoff Procedure was chosen. Second, panel members engaged in a half-hour discussion of what minimum competency means in the insurance field. They were asked to envision situations and then discuss what the minimally competent insurance person would need to know in such situations to protect the public welfare. The test development specialist listed the attributes of such minimally competent individuals on a black-board. Third, the panel members were asked to review the instructions for the question rating form, which is presented intact in Figure 4.4.2. The panel members had received the instructions in the packet mailed to them, and they were asked to review to refresh their memories and to ask any questions. Fourth, the panel members were presented with eight sample questions and asked to apply the rating procedure. They used the same seven point scale that they were to use for the 100 questions to be assessed. (See the Appendix for the Question Rating Form.) This scale is as follows:

Estimated Percentage of Minimally Knowledgeable Individuals Who Know Answer to Questions							
5	20	40	60	75	90	95	DNK

Where DNK stands for "do not know."

Some comments about choice of scale points are important to make. First, the options are centered around 60 since the average percent correct on Part 1's of the four tests in the past has centered around 60%. Two, while the other options are then spaced on either side of 60, the 70 scale point was avoided because this is typically

### Instructions for Question Rating Form

Your task is to make judgements about the difficulty of individual test questions for minimally knowledgeable persons in the lines of insurance covered by the tests you will be reviewing. You will be asked to draw upon your experience to construct a hypothetical group of persons, each of whom, in your judgment, has the *minimum* amount of knowledge to assure the public that only competent individuals are licensed to sell insurance. This study is concerned with individuals who are *just entering* the profession of insurance and have little if any previous work experience. Within the seven states in the Program, only Massachusetts has educational prerequisites for applying for a license and most candidates have either studied for the tests independently or participated in a company sponsored training course.

Your judgments about the test questions are to be made with reference to your conception of a group of minimally knowledgeable individuals as described in the preceding paragraph. As you read each test question and its answer, think of this group. Judge what percentage of the persons in the group would be able to identify or arrive at the answer to the question. If there were 100 minimally knowledgeable individuals, how many of them would know the answer?

When you have made your estimate, circle the percentage on the Rating Form that is closest to your estimate. Before you circle the percentage, please make sure that the number that identifies the question on the form is the same as the number that identifies the question in your question booklet.

If you feel that your experience provides you with no basis whatsoever for making a judgment about one of the questions, you may circle "DNK" (for "Do Not Know"). The DNK category is not to be used simply because you have difficulty in deciding upon a percentage estimate; you are to make a decision even if it is a difficult one. The DNK category is to be used *only* when you have *no* basis for making any judgment.

In making your judgments, you are not to be concerned about how many questions you are assigning to the various percentage categories. It is your responsibility to apply your best judgment in evaluating each question individually.

Figure 4.4.2.

a cut-off score on state insurance tests, and we wanted to avoid the possibility of over-selection of that point simply because of familiarity.

Returning to the discussion of the sample questions, after the panel had responded, they were asked to raise their hands when the scale point that they had chosen was called, and a histogram was built. Members who made choices widely disparate from the group average were asked to explain why they had done so. Some discussion usually ensued. Finally, the members were provided with the item difficulty values for the sample items, which were taken from old test forms. While the item difficulty, or proportion correct value, does not directly translate into "minimally competent" performance, it none-the-less gave the panel some indication about how the item performed in the past. Further, research done (Lorge & Kruglov, 1952; Lorge & Diamond, 1954) indicates that judges tend to overestimate the difficulty of easy questions and underestimate the difficulty of hard items. This was true of the panel members for this study, particularly in reference to the more difficult sample items. In a few instances, in fact, the proportion of minimally competent individuals who would know the answer closely coincided with the actual item difficulty. Some actual performance data helped to point out the fact that certain of the members were overestimating the capabilities of the minimally competent group.

#### 4.4.5 Panel Tasks—Content Rating Form

Prior to the content rating task, and after a discussion of the question rating task, the test development specialist discussed the content specifications or outline and how she had keyed each of the items to be judged to the content specifications. The panel members were then asked to reread the instructions for the task, which are presented intact in Figure 4.4.3, and final questions were entertained. A brief explanation was also offered about how the estimates they were asked to make about the number of questions in each major content area that the minimally competent person would answer correctly would give an alternate procedure for setting a cut-score. The four content raising forms, one for each test, are presented in the Appendix.

#### 4.4.6 Results—Cut-off Scores

Tables 4.2.4 thru 4.4.6 present the results of the question rating form and the section of the content rating form that deals with setting a cut score. Table 4.4.7 presents a comparison of the cut-score arrived at by each of the procedures. Some comments can now be made about the cut-off scores arrived at:

1. For all four tests, the cut-off score generated from the content rating was higher than that generated from the question rating form. The differences between cut-off scores for the two procedures ranged from 1 to 1.6 questions across the four tests.
2. For the question rating task, the half-panels that assessed the life and accident/health exams were very similar in their assessments. The half-panels that assessed the property and casualty tests tended to be less consistent in their ratings.

### Instructions for Content Rating Form

Your task is to examine the description of the test content (Content Outline) of each of the examinations in relation to each test question and to ascertain whether or not these content areas and questions are appropriate for a minimum competency test.

In making your judgment, review the content outline of each test, paying particular attention to the description of the levels of difficulty at which the questions are developed for each of the sections of the Outline. Each of the questions was developed following the guidelines set out in the Content Outlines and the instructions presented in the booklet entitled "Guide to Question Writing."

When you have evaluated the question in relation to the Content Outline, circle your decision on the Rating Form. Before you circle your choice, please make sure that the number that identifies the question on the form is the same as the number that identifies the question in your question booklet.

After evaluating each question and making your response on the rating form, answer the questions about your estimates of success on the major sections of the test.

If you feel that your experience provides you with no basis whatsoever for making a judgment about one of the questions, you may circle "DNK" (for "Do Not Know"). The DNK category is not to be used simply because you have difficulty in deciding upon a percentage estimate; you are to make a decision even if it is a difficult one. The DNK category is to be used *only* when you have *no* basis for making any judgment.

Figure 4.4.3.



Table 4.4.2

Estimates of Average Number of Answers Known by the  
Minimally Knowledgeable Applicant Group  
(Question Rating Form)

Test	Panel 1		Panel 2		Total Panel	
	N	# Known	N	# Known	N	# Known
Life	9	33.6	8	33.5	17	33.6
Accident/Health	7	34.1	8	34.2	15	34.1
Property	9	34.6	9	31.6	18	33.1
Casualty	9	33.3	9	30.9	18	32.0

Table 4.4.3

Number of Questions a Minimally Competent Person  
Would Answer Correctly  
(Content Rating Form)

Life Test

Test Content	Panel 1	Panel 2
A—Basic Principles and Concepts (10 questions)	7.4	7.6
B—Life Insurance Provisions (20 questions)	12.8	14.4
C—Kinds of Insurance and Annuities (20 questions)	14.6	13.6
TOTALS	34.8	35.6
TOTAL PANEL (Average)	35.2	

Table 4.4.4

Number of Questions a Minimally Competent Person  
Would Answer Correctly  
(Content Rating Form)

Accident and Health Test

Test Content	Panel 1	Panel 2
A—Basic Principles and Concepts (10 questions)	7.0	7.5
B—Individual Accident and Health Provisions (20 questions)	13.0	13.5
C—Types of Coverage (15 questions)	11.3	10.5
D—Types of Contracts (5 questions)	3.5	4.4
TOTALS	34.5	35.9
TOTAL PANEL (Average)	35.3	

Table 4.4.5

Number of Questions a Minimally Competent Person  
Would Answer Correctly  
(Content Rating Form)

Property Test

Test Content	Panel 1	Panel 2
A—Basic Principles and Concepts (17 questions)	12.1	11.8
B—Standard Fire Policy (12 questions)	8.3	8.6
C—Forms and Endorsements (10 questions)	6.0	6.4
D—Package Policies (10 questions)	7.0	7.0
E—Flood Insurance (1 question)	.4	.9
TOTALS	33.9	34.7
TOTAL PANEL AVERAGE		34.3

Table 4.4.6

Number of Questions a Minimally Competent Person  
Would Answer Correctly  
(Content Rating Form)

Casualty Test

Test Content	Panel 1	Panel 2
A—Basic Principles and Concepts (15 questions)	10.5	10.1
B—Basic Concepts of Auto Insurance (15 questions)	9.8	9.9
C—General Liability Contracts (13 questions)	7.9	7.3
D—Crime Insurance (5 questions)	4.1	3.2
E—General Principles of Suretyship (2 questions)	1.8	1.4
TOTALS	34.0	31.9
TOTAL PANEL (Average)	33.0	

Table 4.4.7

A Comparison of the Cut-off Scores For  
the Two Procedures Used

Test	Question Rating	Content Rating
Life	33.6	35.2
Accident/Health	34.1	35.3
Property	33.1	34.3
Casualty	32.0	33.0

3. For the content rating form, there are no discernible trends between the panels. Average tended to fairly closely coincide, with the largest difference being between half-panel 1 (34.0) and half-panel 2 (31.9) on the casualty test.

The four tests assessed by the panels were administered to candidates in the seven states that are members of MILP on Saturday, November 11. The tests were equated back to base tests, using common item equating, on November 14-15, and then placed on a scale ranging from 50 to 100. The raw and scaled cut-offs for the four tests are presented below. The raw cut-offs indicated were established using the Modified-Angoff Technique.

<u>Test</u>	<u>Raw Cut-off</u>	<u>Scaled Cut-off</u>
Life	33	80
Accident-Health	34	77
Property	33	77
Casualty	32	79

The scaled scores corresponding to the raw score cut-offs established for each of these tests will be presented to the Commissioner to aid in the setting of state scaled cut-off scores. In the past, this scaled cut-off has been set more or less arbitrarily at either 70 or 75. The present data seems to suggest that scaled cut-offs of higher than 75 are in line. While certainly not a rationale for raising a cut-score, many of the Commissioners have voiced concern that too many examinees are passing the test in their

states. The results of this study should afford the Commissioners a more defensible ground for setting a cut-off score, and as an aside, solve the problems of over-certification.

#### 4.4.7 Results—Content Validity

While content validation procedures are not the major focus of this chapter of the dissertation, the results of the MILP content validation are presented because content validation is necessary before a cut-score can be established. Tables 4.4.8 thru 4.4.11 present the number of questions judged content appropriate by 75% or greater of the judges. Tables 4.4.12 thru 4.4.15 present the panel responses regarding the appropriateness of the content area for a minimum competency test. Little more can be said about the items themselves because they are secure, and hence, can't be reproduced in this document. The information was provided to the test development staff and the few questions judged content inappropriate will be subsequently either revamped or removed from the test. The areas judged content inappropriate will also be closely assessed.

#### 4.4.8 Comments on Cut-Score Procedures

In applying the procedures for setting cut-off scores, a number of problems or situations deserving comment arose. What follows are some observations that may prove useful to anyone implementing either the procedures discussed here, or generally, any



Table 4.4.8

Number of Questions Judged Content Appropriate  
by 75% or Greater of the Judges

Life Test

Test Content	Panel 1	Panel 2	Total Panel
A—Principles and Concepts (10 questions)	10	10	10
B—Life Insurance Provisions (20 questions)	19	18	20
C—Kinds of Insurance and Annuities (20 questions)	16	17	17

Table 4.4.9

Number of Questions Judged Content Appropriate  
by 75% or Greater of the Judges

Accident and Health Test

Test Content	Panel 1	Panel 2	Total Panel
A—Basic Principles and Concepts (10 questions)	10	10	10
B—Individual Accident and Health Provision (20 questions)	18	20	20
C—Types of Coverage (15 questions)	12	14	12
D—Types of Contracts (5 questions)	5	5	5

Table 4.4.10

Number of Questions Judged Content Appropriate  
by 75% or Greater of the Judges

Property Test

Test Content	Panel 1	Panel 2	Total Panel
A—Basic Principles and Concepts (17 questions)	16	16	16
B—Standard Fire Policy (12 questions)	10	10	10
C—Forms and Endorsements (10 questions)	10	10	10
D—Package Policies (10 questions)	10	10	10
E—Flood Insurance (1 question)	1	1	1

Table 4.4.11

Number of Questions Judged Content Appropriate  
by 75% or Greater of the Judges

Casualty Test

Test Content	Panel 1	Panel 2	Total Panel
A—Basic Principles and Concepts (15 questions)	15	15	15
B—Basic Concepts of Auto Insurance (15 questions)	15	15	15
C—General Liability Contracts (13 questions)	13	12	12
D—Crime Insurance	5	3	5
E—General Principles of Suretyship (2 questions)	2	2	2

Table 4.4.12

Panel Responses Regarding the Appropriateness  
of Content AreasLife Test

Test Content	Panel 1			Panel 2			Total Panel		
	Yes	No	% Yes	Yes	No	% Yes	Yes	No	% Yes
A—Basic Principles and Concepts (10 questions)	8	0	100	8	0	100	16	0	100
B—Life Insurance Provisions (20 questions)	7	1	87.5	8	0	100	15	1	93.75
C—Kinds of Insurance and Annuities (20 questions)	8	0	100	8	0	100	16	0	100

Table 4.4.13

Panel Responses Regarding the Appropriateness  
of Content AreasAccident and Health Test

Test Content	Panel 1			Panel 2			Total Panel		
	Yes	No	% Yes	Yes	No	% Yes	Yes	No	% Yes
A—Basic Principles and Concepts (10 questions)	5	0	100	7	1	87.5	12	1	92.3
B—Individual Accident and Health Provisions (20 questions)	4	1	80	8	0	100	12	1	92.3
C—Types of Coverage (15 questions)	5	0	100	6	2	75	11	2	84.6
D—Types of Contracts (5 questions)	4	0	100	6	2	75	10	2	83.3

Table 4.4.14

Panel Responses Regarding the Appropriateness  
of Content AreasProperty Test

Test Content	Panel 1			Panel 2			Total Panel		
	Yes	No	% Yes	Yes	No	% Yes	Yes	No	% Yes
A—Basic Principles and Concepts (17 questions)	9	0	100	9	0	100	18	0	100
B—Standard Fire Policy (12 questions)	9	0	100	9	0	100	18	0	100
C—Forms and Endorsements (10 questions)	9	0	100	8	1	88.8	17	1	94.4
D—Package Policies (10 questions)	9	0	100	8	1	88.8	17	1	94.4
E—Flood Insurance (1 question)	9	0	100	8	1	88.8	17	1	94.4

Table 4.4.15

Panel Responses Regarding the Appropriateness  
of Content AreasCasualty Test

Test Content	Panel 1			Panel 2			Total Panel		
	Yes	No	% Yes	Yes	No	% Yes	Yes	No	% Yes
A—Basic Principles and Concepts (15 questions)	8	0	100	9	0	100	17	0	100
B—Basic Concepts of Auto Insurance (15 questions)	8	0	100	9	0	100	17	0	100
C—General Liability Contracts (13 questions)	8	0	100	9	0	100	17	0	100
D—Crime Insurance (5 questions)	8	0	100	9	0	100	17	0	100
E—General Principles of Suretyship (2 questions)	8	0	100	9	0	100	17	0	100

judgmental procedure for setting a cut-off score. The observations are as follows:

1. A twenty minute discussion on standards is not sufficient to introduce a group of judges to the general need for cut-off scores. This author had to spend about five minutes alone simply explaining how a cut-score worked in reference to pass-fail decisions.
2. A general discussion of minimum competency in the area being assessed must be allotted a greater amount of time than done in this study. We were aware of the problem, but simple logistics dictated that such time had to be kept short. Zieky and Livingston (1977) suggest two to three hours be spent in reaching a definition of minimally acceptable performance.
3. Some sort of discussion about how to consider the problem of guessing should take place. For instance, should the judges build into their estimates of the percentage of minimally knowledgeable individuals who would get the answer correct the fact that certain individuals will guess the question correctly. We tried to circumvent the problem by instead wording the question rating form as the "estimated percentage of minimally knowledgeable individuals who know the answer to the question," but confusion still arose, and had to be clarified. A better ploy would be to discuss and clarify the problem beforehand.
4. A careful clarification between the statement "Judge what percentage of the persons in the group would be able to identify or arrive at the answer to the question" (taken from instructions) and "Judge what percentage of the persons in the group should be able to identify or arrive at the answer to the question" is essential. A great deal of confusion existed in regard to this point, and only through careful verbal clarification were we able to assure that "would be able" was to be used, rather than the evaluative "should be able."
5. The use of sample questions and related normative data was a decided plus in this study. Feedback from participants indicated that the practice session clarified both their task and their notion of minimum competence in insurance.
6. Supplemental data that may be of use in the future should the need arise for the setting of further cut-points involves the performance of students whose scores are adjacent to the cut-off. This would provide the panel with a better indication of how the borderline group on the test performed on each of the sample questions.

7. Finally, it would be advantageous to combine more than one well-established method for arriving at a cut-off score in the procedures being used. This would serve both a research and a practical function. From a research prospective, little has been done to date comparing methods. The Andrew and Hecht (1976) study has been the only one to appear in the literature to date. From a practical viewpoint, a concern about the lack of validity of judgmental procedures could be partially alleviated if two procedures led to the same (or perhaps with a small margin of difference) cut-score. Of course, one has to be prepared with what to do with widely disparate cut-scores. For this study, that was to prove to be the ultimate concern, and hence only one rigorous method was used. The Project Director felt, and this author agrees, that the difficulties involved in explaining differences would be so great that any impact the study would have on the Insurance Commissioner's judgments about a standard would be negated.

#### 4.5 Conclusion

In Chapter One of this dissertation, two objectives were specified to guide the research on cut-off scores that was presented in this chapter. These objectives were:

1. The organization of the available methods for setting cut-off scores in a useful form for practitioners.
2. The presentation of guidelines and implementation strategies to aid individuals in answering the following questions: "How can the 'best' method for use in a prototypical situation be selected?" and "How should the chosen method be implemented?"

In reference to the first objective, the first half of Chapter Four involved an organization of available methods for setting cut-scores for criterion-referenced tests. Methods that were not suitable were rejected and the remaining methods were organized into three sets: judgmental models, empirical models, and combination models. Each of the methods were then presented along with relevant examples and discussion.

In reference to the second objective, the prototypical situation chosen was minimum competency testing. The three sets of models were considered for application to minimum competency testing, and it was found that only certain of the models were applicable. These models were compared and a final choice of two models most suitable for minimum competency tests, the Contrasting Groups Method and Modified-Angoff Technique, was presented. Finally, the Modified-Angoff Technique was applied to four tests in the Insurance area for the purpose of setting multistate cut-points. A discussion of this experience and suggestions for future use of the Modified-Angoff Technique was presented. A discussion of necessary further research in the area is presented in Chapter Five.

## C H A P T E R V

### CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

Hambleton, Swaminathan, Algina and Coulson (1978), in their review of the criterion-referenced testing field, offered seven suggestions for further research. Two of their suggestions, the need for further research on the topic of test length and reliability, and the need for better organization of cut-score methods, along with useful implementation strategies, have been investigated in this dissertation. The third area investigated in this dissertation, the establishment of guidelines for evaluating criterion-referenced tests and test manuals, was discussed in several places in the Hambleton et al., review. Hence, the timeliness of the research reported in this dissertation appears evident.

Popham (1978), Hambleton et al. (1978), and Hambleton and Eignor (1978a) have done a thorough job of suggesting topics for further research. Therefore, the comments to be made in this chapter will address specific research that could be done in each of the three topic areas covered in this study.

In reference to Guidelines for Evaluating Criterion-Referenced Tests and Test Manuals (Chapter II), one area of further research is immediately evident. While the guidelines are presently in a form that is understandable to the practitioner, they are not at an



operational level. Also, it would be useful if the ratings attached to each guideline contained examples of what is acceptable and unacceptable. Such additions to the guidelines will improve their usefulness.

Two additional activities are desirable. One, the guidelines need extensive review by educational measurement specialists and other groups with an interest in the ways tests are selected and used. This will certainly lead to an explication of any relevant concern left out of the guidelines. Two, hopefully the work presented here will operate as a catalyst for both further discussion and other sets of guidelines. A discussion and subsequent merger of independently developed guidelines would certainly be of use to the criterion-referenced test user.

The chapter addressing the relationship of criterion-referenced test length to reliability and validity is a good initial start in developing technical materials that are useful for the practitioner. Past work done in the area (Novick & Lewis, 1974; Wilcox, 1976) tends to be conceptually difficult for the practitioner to understand. Hopefully this is not the case with the development offered in this dissertation.

Two lines of research appear to be necessary. For one, it is soon going to be necessary for educational measurement specialists to reach a consensus about what constitutes a suitable level of reliability for a criterion-referenced test. Suitable guidelines have existed for some time for norm-referenced tests. For instance, one wants the reliability of norm-referenced achievement tests to be above .90 and aptitude tests to be at least .80 (Stanley & Hopkins,

1972). No such guideline exists for criterion-referenced tests. Of course, much of the direction for such guidelines will come from empirical research, which is the second sort of necessary research. What is the nature of the various reliability and validity indices with real data? How close do the indices based on real data come to the indices offered in this dissertation? These and other questions need to be addressed in order to offer criterion-referenced test constructors and evaluators some concrete decision procedures about the reliability of their tests.

The third area of research reported in this dissertation is frequently discussed because of the minimum competency testing movement in today's schools. The pros and cons of setting standards have been debated at many levels, most recently in the Journal of Educational Measurement (Vol. 15, No. 4, Winter 1978). One thing is certain; the minimum competency testing movement is a reality, and hence, cut-score methods, good or bad, are going to be used. The work presented here should be helpful in pointing out which methods for setting cut-scores are useful in minimum competency testing programs.

There are at least three topics requiring further research. One, and perhaps most important, there needs to be further articulation of implementation strategies for setting cut-offs for the variety of uses that exist for criterion-referenced tests. This dissertation addresses the setting of cut-scores for licensing and certification minimum competency tests. The only other work done to date that involves implementation procedures is by Zieky and

Livingston (1977) and Popham (1978), and both address the classroom instructional setting. Examples of other areas where guidelines for both choosing and implementing standard-setting methods are essential include program evaluation and performance testing.

Second, more research needs to be done on methods that require the input of judges. There is a body of literature that exists on group dynamics and group decision-making procedures that is relevant for those cut-score methods that require judgmental input. For instance, is the Delphi Method potentially suitable for use in setting standards? In reference to this area of research, a group of colleagues at the University of Massachusetts are presently beginning investigation in the area.

Third, there needs to be considerably more study of the term "minimally competent" because if the term is better understood, it may be possible to link existing standard-setting methods to the intended meaning or meanings of the term and thereby greatly facilitate the selection of a standard-setting method or the development of new methods. This is critical for the minimum competency testing movement, and also for those judgmental procedures that require a definition of minimum competence for operation.

## R E F E R E N C E S

- Andrew, B. J., & Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement*, 1976, 36, 45-50.
- Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement*. Washington, D.C.: American Council on Education, 1971.
- Berk, R. A. Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, 1976, 45, 4-9.
- Block, J. H. (Ed.) *Mastery learning: Theory and practice*. New York: Holt, Rinehart and Winston, 1971.
- Block, J. H. Student learning and the setting of mastery performance standards. *Educational Horizons*, 1972, 50, 183-190.
- Bloom, B. S. Learning for mastery. *Evaluation Comment*, 1968, 1(2).
- Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. *Journal of Educational Measurement*, 1977, 14, 277-289.
- Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. *Psychometrika*, 1977, 42, 609-625.
- Brown, F. G. *Principles of educational and psychological testing*. (2nd ed.) New York: Holt, Rinehart and Winston, 1976.
- Burton, N. Societal standards. *Journal of Educational Measurement*, 1978, 15, 263-271.
- 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.
- Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, 20, 37-46.

- Conaway, L. E. Discussant comments: Setting performance standards based on limited research. *Florida Journal of Educational Research*, 1976, 18, 35-36.
- Conaway, L. E. Setting standards in competency-based education: Some current practices and concerns. Paper presented at the annual meeting of NCME, New York, 1977.
- Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement*. (2nd ed.) Washington: American Council on Education, 1971.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972.
- CSE Secondary School Test Evaluations*. Los Angeles: Center for the Study of Evaluation, UCLA, 1974.
- Ebel, R. L. *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall, 1972.
- Educational Testing Service. Report on a study of the use of the National Teachers Examination by the State of South Carolina. Princeton, NJ: Educational Testing Service, 1976.
- Emrick, J. A. An evaluation model for mastery testing. *Journal of Educational Measurement*, 1971, 8, 321-326.
- Fhaner, S. Item sampling and decision-making in achievement testing. *British Journal of Mathematical and Statistical Psychology*, 1974, 27, 172-175.
- Flanagan, J. C. Functional education for the seventies. *Phi Delta 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., Shaycoft, M. F., Orr, D. B., Goldberg, I., & Neyman, C. A., Jr. *The American high school student*. Cooperative Research Project No. 635, U.S. Office of Education) Pittsburgh: American Institute for Research and University of Pittsburgh, 1964.

- Gibbons, M. What is individualized instruction? *Interchange*, 1970, 1, 28-52.
- Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist*, 1963, 18, 519-521.
- Glaser, R. Adapting the elementary school curriculum to individual performance. In *Proceeding of the 1967 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1968.
- Glaser, R. Evaluation of instruction and changing educational models. In M. C. Wittrock and D. E. Wiley (Eds.), *The evaluation of instruction*. New York: Holt, Rinehart and Winston, 1970.
- 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.
- Glass, G. V. Standards and criteria. *Journal of Educational Measurement*, 1978, 15, 237-261. (a)
- Glass, G. V. Minimum competence and incompetence in Florida. *Phi Delta Kappan*, 1978, 59, No. 9 (May), 602-605. (b)
- Gronlund, N. E. *Individualizing classroom instruction*. New York: Macmillan Publishing Co., 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. Validation of criterion-referenced test score interpretations. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands, 1977.
- Hambleton, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, 1978, 15, 277-290.
- Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. *Laboratory of Psychometric and Evaluative Research Report No. 70*. Amherst, MA: School of Education, University of Massachusetts, 1978. (a)

- Hambleton, R. K., & Eignor, D. R. Guidelines for evaluating criterion-referenced tests and test manuals. *Journal of Educational Measurement*, 1978, 15, 321-327.
- Hambleton, R. K., & Eignor, D. R. Competency test development, validation, and standard-setting. Paper presented at the AERA Minimum Competency Testing Conference, Washington, October 12-14, 1978. (c).
- Hambleton, R. K., & Gifford, J. A. Development and use of criterion-referenced tests to evaluate program effectiveness. *Laboratory of Psychometric and Evaluative Research Report No. 52*. Amherst, MA: School of Education, University of Massachusetts, 1977.
- 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., & Rovinelli, R. A Fortran IV program for generating examinee response data from logistic test models. *Behavioral Science*, 1973, 17, 73-74.
- Hambleton, R. K., Swaminathan, H., & Algina, J. Some contributions to the theory and practice of criterion-referenced testing. In D.N.M. de Gruijter, & L. J. Th. van der Kamp (Eds.), *Advances in psychological and educational measurement*. New York: Wiley, 1976.
- Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. Symposium presented at the annual meeting of AERA, Washington, D.C., 1975.
- Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, 48, 1-47.
- Harris, C. W., Alkin, M. C., & Popham, W. J. *Problems in criterion-referenced measurement*. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.
- Heathers, G. Overview of innovations in organization for learning. *Interchange*, 1972, 3, 47-68.

- Hively, E., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. *Domain-referenced curriculum evaluation: A technical handbook and a case study from the Minnemast Project*. CSE monograph series in evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California, 1973.
- Hsu, G. Mastery test reliability: Relation to the size of the indifference zone and number of test items. Unpublished manuscript, 1977.
- Huynh, H. On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 1963, 13, 253-264. (a)
- Huynh, H. Statistical consideration of mastery scores. *Psychometrika*, 1976, 41, 65-78. (b)
- Huynh, H., & Perney, J. Determination of mastery scores when instructional units are linearly related. *Educational and Psychological Measurement*, in press.
- Jaeger, R. M. Measurement consequences of selected standard-setting models. *Florida Journal of Educational Research*, 1976, 18, 22-27.
- Jaeger, R. M. A proposal for setting a standard on the North Carolina High School Competency Test. Paper presented at the 1978 spring meeting of the North Carolina Association for Research in Education, Chapel Hill, 1978.
- Kriewall, T. E. Aspects and applications of criterion-referenced tests. Paper presented at the annual meeting of AERA, Chicago, 1972.
- Linn, R. L. Issues of validity in measurement for competency-based programs. Paper presented at the meeting of the National Council on Measurement in Education, New York, 1977.
- Livingston, S. A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, 9, 13-26.



- Livingston, S. A. A utility based approach to the evaluation of pass/fail testing decision procedures. COPA Research Report. Princeton, NJ: Educational Testing Service, 1975.
- Livingston, S. A. Choosing minimum passing scores by stochastic approximation techniques. *Report No. COPA-76-02*. Princeton, NJ: Center for Occupational and Professional Assessment, Educational Testing Service, 1976.
- Livingston, S. A. Assessing the reliability of tests used to make pass/fail decisions. COPA Research Report. Princeton, NJ: Educational Testing Service, 1978.
- Lord, F. M. A strong true-score theory, with applications. *Psychometrika*, 1965, 30, 239-270.
- Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.
- Lorge, I., & Diamond, L. The value of information to good and poor judges of item difficulty. *Educational and Psychological Measurement*, 1954, 14, 29-33.
- Lorge, I., & Kruglov, L. A suggested technique for the improvement of difficulty prediction of test items. *Educational and Psychological Measurement*, 1952, 12, 554-561.
- Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 1977, 2, 99-120.
- Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research*, 1976, 46, 133-158.
- Messick, S. A. 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, California: McCutchan Publishing Co., 1974.
- Nassif, P. M. Standard-setting for criterion-referenced teacher licensing tests. Paper presented at the annual meeting of NCME, Toronto, 1978.

- Nedelsky, L. Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 1954, 14, 3-19.
- Novick, M. R., & Jackson, P. H. *Statistical methods for educational and psychological research*. New York: McGraw-Hill, 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.
- Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportion in m groups. *Psychometrika*, 1968, 38, 95-104.
- Popham, W. J. (Ed.). *Criterion-referenced measurement: An introduction*. Englewood Cliffs, NJ: Educational Technology Publications, 1971.
- Popham, W. J. *Educational evaluation*. Englewood Cliffs, NJ: Prentice-Hall, 1975.
- Popham, W. J. *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall, 1978. (a)
- Popham, W. J. Setting performance standards. Los Angeles: Instructional Objectives Exchange, 1978. (b)
- Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, 6, 1-9.
- Roudabush, G. E. Models for a beginning theory of criterion-referenced tests. Paper presented at the annual meeting of NCME, Chicago, 1974.
- Schoon, C. G., Gullion, C. M., & Ferrara, P. Credentialing examinations, Bayesian statistics, and the determination of passing points. Paper presented at the annual meeting of APA, Toronto, 1978.
- Shepard, L. A. Setting standards and living with them. *Florida Journal of Educational Research*, 1976, 18, 23-32.
- Standards for Educational and Psychological Tests*. Prepared by a joint committee of the American Psychological Association, American Research Association, and the National Council on Measurement in Education. Washington, D.C.: American Psychological Association, 1974.

- Stanley, J. C., & Hopkins, K. D. *Educational and psychological measurement and evaluation*. Englewood Cliffs, NJ: Prentice-Hall, 1972.
- Subkoviak, M. Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 1976, 13, 265-275.
- Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 1974, 11, 263-268.
- Swaminathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement*, 1975, 12, 87-98.
- Swazey, R. W., & Pearlstein, R. B. *Guidebook for developing criterion-referenced tests*. A report prepared for the U.S. Army Research Institute for the Behavioral and Social Sciences. Reston, VA: Applied Science Associates, August 1975.
- Van der Linden, W. J., & Mellenbergh, G. J. Optimal cutting scores using a linear loss function. *Applied Psychological Measurement*, 1977, 1, 593-599.
- Walker, C. B. *Standards for evaluating criterion-referenced tests*. Los Angeles: Center for the Study of Evaluation, UCLA, 1977.
- Washburne, C. W. Educational measurements as a key to individualizing instruction and promotions. *Journal of Educational Research*, 1922, 5, 195-206.
- Wilcox, R. A note on the length and passing score of a mastery test. *Journal of Educational Statistics*, 1976, 1, 359-364.
- Wilcox, R. R. Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. *Journal of Educational Statistics*, 1977, 2, 289-307.
- Wilcox, R. R. Estimating true score in the compound binomial error model. *Psychometrika*, 1978, 43, 245-258.
- Zieky, M. J., & Livingston, S. A. *Manual for setting standards on the Basic Skills Assessment Tests*. Princeton, NJ: Educational Testing Service, 1977.

APPENDIX

Appendix One

## Test Length Reliability-Validity Simulation

<u>Table</u>	<u>Distribution</u>	<u>Simulation</u>	<u>Classification Weights</u>
1	One	User Prior	Equal
2	One	User Prior	Unequal
3	Two	User Prior	Equal
4	Two	User Prior	Unequal
5	Three	User Prior	Equal
6	Four	User Prior	Equal
7	Five	Beta Prior	Equal
8	Five	Beta Prior	Unequal
9	Six	Beta Prior	Equal
10	Six	Beta Prior	Unequal
11	Seven	Beta Prior	Equal
12	Eight	DATGEN	Equal
13	Nine	DATGEN	Equal
14	Ten	DATGEN	Equal

Table 1  
Simulation One - All Test Lengths and Cut-Offs Simulated:  
Equal Classification Weights

		Reliability				Validity				
No. of Items	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency
2	1	.855	.094	.889	.872	.630	.275	.130	.185	.181
2	2	.610	.154	.885	.565	.690	.364	.190	.155	.507
4	3	.780	.397	.802	.710	.720	.445	.220	.140	.437
4	4	.615	.225	.835	.551	.680	.383	.180	.160	.547
6	4	.790	.332	.826	.792	.715	.448	.215	.143	.485
6	5	.660	.282	.767	.607	.765	.520	.265	.118	.641
8	6	.700	.286	.792	.707	.755	.509	.255	.123	.596
8	7	.685	.370	.763	.627	.790	.584	.290	.105	.789
10	7	.830	.496	.820	.781	.725	.468	.225	.138	.550
10	8	.745	.442	.780	.712	.805	.607	.305	.098	.768
10	9	.715	.425	.772	.669	.730	.478	.230	.135	.745
15	11	.850	.636	.829	.787	.810	.636	.310	.095	.717
15	12	.815	.614	.806	.734	.825	.644	.325	.088	.872
15	13	.740	.478	.778	.724	.800	.613	.300	.100	.844
20	16	.760	.504	.804	.756	.840	.674	.340	.080	.901
20	18	.795	.538	.801	.759	.740	.551	.240	.130	.786
40	32	.795	.580	.823	.798	.830	.655	.330	.083	.887

Table 2  
Simulation One - All Test Lengths and Cut-Offs Simulated:  
Unequal Classification Weights

No. of Items	Cut-Off	Reliability					Validity				
		Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency	Weighted Efficiency
2	1	.905	.249	.889	.923	.625	.272	.016	.278	.072	.156
2	2	.640	.210	.889	.564	.705	.395	.109	.180	.591	.548
4	3	.755	.396	.799	.692	.750	.502	.146	.171	.549	.466
4	4	.685	.367	.843	.582	.720	.458	.134	.137	.685	.714
6	4	.785	.323	.808	.787	.710	.450	.102	.212	.438	.314
8	6	.805	.533	.796	.716	.780	.563	.175	.150	.717	.664
8	7	.675	.349	.758	.666	.815	.625	.217	.103	.807	.807
10	7	.810	.430	.818	.792	.750	.516	.143	.178	.571	.475
10	8	.780	.542	.786	.691	.840	.674	.239	.094	.811	.806
10	9	.760	.508	.760	.680	.765	.567	.179	.103	.783	.835
15	11	.800	.533	.813	.793	.780	.563	.175	.150	.719	.662
15	12	.710	.399	.801	.735	.825	.644	.225	.101	.848	.838
15	13	.715	.428	.794	.713	.820	.647	.226	.088	.848	.863
20	16	.775	.542	.806	.766	.865	.726	.264	.075	.880	.781
20	18	.755	.474	.806	.759	.780	.629	.197	.084	.805	.875
40	32	.790	.570	.835	.813	.860	.715	.258	.081	.879	.871

Table 3

Simulation Two - All Test Lengths and Cut-Offs Simulated:  
Equal Classification Weights

No. of Items	Reliability			Validity						
	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency
2	1	.815	.106	.848	.821	.320	.185	-.185	.340	-.344
2	2	.645	.288	.848	.546	.635	.300	.135	.183	.481
4	3	.610	.188	.755	.572	.540	.255	.040	.230	.246
4	4	.680	.218	.819	.683	.795	.424	.295	.103	.760
6	4	.685	.292	.760	.640	.535	.319	.035	.233	.291
6	5	.675	.344	.756	.630	.735	.480	.235	.133	.672
8	6	.665	.329	.750	.675	.650	.405	.150	.175	.595
8	7	.675	.245	.759	.659	.785	.444	.285	.108	.842
10	7	.740	.468	.768	.685	.560	.306	.060	.220	.434
10	8	.735	.463	.754	.690	.690	.416	.190	.155	.663
10	9	.765	.331	.788	.754	.830	.499	.330	.085	.828
15	11	.740	.482	.786	.735	.670	.455	.170	.165	.665
15	12	.745	.469	.772	.712	.755	.531	.255	.123	.784
15	13	.795	.398	.795	.783	.855	.551	.355	.073	.893
20	16	.755	.476	.807	.755	.830	.641	.330	.085	.850
20	18	.860	.479	.862	.856	.885	.615	.385	.058	.925
40	32	.830	.584	.828	.836	.875	.663	.375	.063	.945



Table 4

Simulation Two - All Test Lengths and Cut-Offs Simulated:  
Unequal Weights

No. of Items	Cut-Off	Reliability				Validity					
		Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency	Weighted Efficiency
2	1	.800	.195	.856	.809	.340	.150	-.359	.491	-.249	-.319
2	2	.535	.073	.842	.527	.600	.280	-.095	.285	.386	.368
4	3	.660	.258	.769	.621	.515	.266	-.183	.356	.287	.253
4	4	.690	.255	.805	.675	.750	.342	.060	.158	.698	.720
6	4	.715	.305	.767	.677	.485	.277	-.214	.383	.203	.155
6	5	.695	.381	.755	.665	.715	.430	.019	.203	.639	.629
8	6	.700	.398	.752	.681	.650	.405	-.048	.257	.615	.597
8	7	.715	.342	.783	.680	.835	.603	.138	.114	.836	.833
10	7	.725	.435	.778	.681	.620	.409	-.079	.283	.471	.443
10	8	.670	.325	.745	.684	.725	.497	.026	.203	.678	.662
10	9	.805	.385	.805	.763	.850	.540	.159	.086	.880	.898
15	11	.760	.521	.786	.717	.710	.481	.011	.214	.695	.676
15	12	.700	.345	.783	.714	.805	.557	.108	.137	.796	.791
15	13	.805	.397	.827	.798	.850	.523	.160	.083	.899	.915
20	16	.710	.367	.802	.772	.815	.596	.117	.133	.817	.811
20	18	.870	.493	.860	.850	.895	.658	.203	.054	.934	.954
40	32	.800	.525	.841	.830	.900	.746	.201	.071	.944	.942

Table 5  
Simulation Three - All Test Lengths and Cut-Offs Simulated:  
Equal Classification Weights

No. of Items	Reliability				Validity					
	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency
2	1	.800	.195	.852	.810	.440	.272	-.060	.280	-.067
2	2	.585	.169	.869	.580	.660	.376	.160	.170	.555
4	3	.695	.375	.760	.615	.665	.435	.165	.168	.544
4	4	.695	.263	.838	.683	.765	.427	.265	.118	.804
6	4	.775	.501	.785	.718	.600	.414	.100	.200	.454
6	5	.725	.440	.760	.670	.725	.459	.225	.138	.733
8	6	.685	.370	.766	.691	.720	.500	.220	.140	.675
8	7	.745	.408	.792	.727	.830	.583	.330	.085	.879
10	7	.705	.401	.789	.719	.680	.463	.180	.160	.619
10	8	.760	.514	.776	.721	.775	.562	.275	.113	.799
10	9	.815	.549	.812	.774	.830	.588	.330	.085	.883
15	11	.800	.499	.818	.786	.745	.581	.245	.128	.774
15	12	.735	.450	.804	.757	.820	.613	.320	.090	.851
15	13	.825	.537	.824	.793	.825	.547	.325	.088	.891
20	16	.800	.580	.813	.793	.875	.736	.375	.063	.918
20	18	.825	.426	.856	.840	.845	.592	.345	.078	.941
40	32	.830	.616	.855	.842	.890	.756	.390	.055	.928

Table 6

Simulation Four - All Test Lengths and Cut-Offs Simulated:  
Equal Classification Weights

No. of Items	Cut-Off	Reliability				Validity				
		Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency
2	1	.920	-.041	.912	.940	.725	.226	.225	.138	.598
2	2	.600	.030	.900	.576	.685	.232	.185	.158	.520
4	3	.770	.128	.816	.817	.715	.214	.215	.143	.574
4	4	.525	.043	.835	.537	.640	.280	.140	.180	.384
6	4	.865	.199	.862	.855	.725	.228	.225	.138	.555
6	5	.695	.149	.774	.732	.720	.308	.220	.140	.652
8	6	.725	.179	.807	.786	.775	.416	.275	.113	.686
8	7	.630	.234	.740	.613	.700	.379	.200	.150	.663
10	7	.840	.157	.845	.848	.795	.476	.295	.103	.710
10	8	.735	.279	.794	.721	.765	.409	.265	.118	.702
10	9	.615	.228	.749	.602	.710	.417	.210	.145	.608
15	11	.855	.213	.843	.867	.755	.345	.255	.123	.671
15	12	.760	.372	.787	.723	.810	.532	.310	.095	.790
15	13	.625	.235	.751	.653	.740	.453	.240	.130	.692
20	16	.770	.403	.784	.741	.815	.532	.315	.093	.797
20	18	.690	.380	.787	.663	.715	.492	.215	.143	.708
40	32	.785	.486	.817	.791	.865	.677	.365	.068	.872



Table 8

Simulation Five - All Test Lengths and Cut-Offs Simulated:  
Unequal Classification Weights

No. of Items	Reliability			Validity							
	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency	Weighted Efficiency
2	1	.855	.049	.895	.871	.630	.309	.018	.274	.065	-.141
2	2	.550	.049	.874	.523	.660	.325	.051	.208	.479	.439
4	3	.690	.127	.785	.699	.640	.347	.016	.253	.308	.182
4	4	.630	.251	.837	.576	.695	.397	.098	.158	.593	.612
6	4	.735	.148	.809	.770	.615	.310	-.009	.276	.298	.165
6	5	.685	.330	.771	.667	.725	.467	.107	.178	.654	.634
8	6	.725	.374	.782	.719	.770	.594	.136	.165	.716	.667
8	7	.710	.419	.754	.604	.765	.529	.145	.135	.733	.741
10	7	.705	.256	.782	.701	.670	.470	.028	.244	.411	.316
10	8	.680	.341	.787	.693	.740	.492	.123	.165	.692	.672
10	9	.710	.410	.769	.657	.765	.544	.164	.113	.806	.853
15	11	.780	.476	.799	.758	.770	.604	.143	.171	.660	.586
15	12	.745	.479	.804	.744	.780	.574	.158	.143	.755	.726
15	13	.760	.508	.767	.679	.785	.577	.174	.103	.795	.832
20	16	.730	.443	.783	.729	.825	.652	.215	.113	.826	.795
20	18	.745	.385	.800	.745	.680	.398	.088	.139	.677	.763
40	32	.805	.608	.835	.803	.855	.713	.234	.090	.875	.863

Table 9  
Simulation Six - All Test Lengths and Cut-Offs Simulated  
Equal Classification Weights

		Reliability				Validity				
No. of Items	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency
2	1	.860	.492	.901	.832	.625	.399	.125	.188	.343
2	2	.805	.609	.878	.634	.830	.673	.330	.085	.783
4	3	.740	.433	.811	.720	.765	.563	.265	.118	.649
4	4	.755	.499	.856	.657	.825	.647	.325	.088	.856
6	4	.825	.556	.844	.809	.755	.579	.255	.123	.636
6	5	.755	.491	.817	.762	.795	.596	.295	.103	.766
8	6	.795	.580	.837	.794	.805	.649	.305	.098	.769
8	7	.780	.559	.826	.771	.825	.651	.325	.088	.867
10	7	.795	.510	.845	.808	.750	.559	.250	.125	.672
10	8	.810	.618	.846	.807	.870	.754	.370	.065	.900
10	9	.840	.668	.841	.760	.890	.778	.390	.055	.931
15	11	.865	.715	.868	.843	.825	.685	.325	.088	.867
15	12	.880	.760	.861	.829	.900	.799	.400	.050	.934
15	13	.820	.629	.829	.807	.865	.725	.365	.068	.922
20	16	.870	.740	.870	.868	.885	.773	.405	.048	.940
20	18	.825	.581	.864	.840	.855	.716	.355	.073	.913
40	32	.920	.839	.923	.918	.935	.870	.435	.033	.976

Table 10  
Simulation Six - All Test Lengths and Cut-Offs Simulated:  
Unequal Classification Weights

No. of Items	Cut-Off	Reliability				Validity					
		Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency	Weighted Efficiency
2	1	.850	.412	.861	.830	.555	.325	-.091	.332	.120	.003
2	2	.720	.439	.867	.625	.760	.540	.136	.158	.670	.645
4	3	.715	.360	.814	.703	.715	.476	.085	.195	.592	.544
4	4	.720	.439	.880	.664	.810	.624	.203	.098	.820	.846
6	4	.825	.596	.835	.814	.805	.651	.181	.143	.697	.647
6	5	.755	.470	.823	.754	.790	.574	.188	.131	.772	.760
8	6	.815	.595	.857	.824	.830	.673	.217	.118	.775	.742
8	7	.775	.543	.828	.781	.865	.729	.223	.084	.896	.900
10	7	.835	.649	.862	.839	.820	.686	.187	.133	.805	.779
10	8	.855	.709	.862	.831	.870	.746	.241	.084	.893	.884
10	9	.810	.614	.846	.805	.840	.688	.232	.077	.883	.906
15	11	.805	.604	.855	.830	.795	.639	.143	.148	.830	.812
15	12	.865	.724	.862	.838	.870	.746	.247	.084	.898	.891
15	13	.880	.780	.848	.846	.900	.801	.283	.051	.956	.961
20	16	.870	.740	.880	.875	.885	.774	.259	.073	.957	.955
20	18	.885	.758	.894	.873	.875	.760	.256	.053	.924	.945
40	32	.920	.840	.919	.914	.960	.920	.336	.021	.989	.991

Table 11  
Simulation Seven - All Test Lengths And Cut-Offs Simulated  
Equal Classification Weights

No. of Items	Reliability				Validity					
	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency
2		.665	.223	.833	.632	.300	.078	-.200	.350	-.264
2	1	.610	.080	.829	.620	.710	.154	.210	.145	.550
4	3	.640	.210	.726	.576	.645	.133	.145	.178	.456
4	4	.845	.077	.850	.871	.925	.392	.425	.038	.900
6	4	.620	.171	.715	.582	.650	.163	.150	.175	.449
6	5	.790	.220	.802	.795	.845	.310	.345	.078	.776
8	6	.710	.173	.777	.724	.810	.147	.310	.095	.727
8	7	.875	-.064	.858	.907	.940	.131	.440	.030	.950
10	7	.720	.293	.774	.723	.765	.126	.265	.118	.747
10	8	.840	.183	.815	.853	.895	.220	.395	.053	.877
10	9	.935	.103	.909	.960	.945	-.016	.445	.028	.959
15	11	.805	.319	.818	.812	.820	.251	.320	.090	.814
15	12	.910	.352	.893	.899	.935	.469	.435	.033	.955
15	13	.935	.100	.923	.967	.965	.361	.465	.018	.988
20	16	.960	.536	.910	.947	.950	.294	.450	.025	.969
20	18	.980	.000	.966	.989	.985	--	.485	.008	1.000
40	32	.985	.524	.942	.961	.970	.568	.470	.015	.982



Table 12  
 All Simulated Test Lengths and Associated Cut-Offs For Simulation Eight:  
 Equal Weights

No. of Items	Cut-Off	Reliability				Validity				
		Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency
2	1	.865	.155	.884	.885	.685	.335	.185	.158	.228
2	2	.615	.234	.827	.509	.650	.338	.150	.175	.394
4	3	.740	.412	.783	.657	.720	.522	.220	.140	.559
4	4	.750	.342	.804	.660	.795	.499	.295	.103	.826
6	4	.750	.472	.758	.615	.590	.364	.090	.205	.400
6	5	.650	.287	.737	.576	.720	.460	.220	.140	.615
8	6	.745	.469	.788	.686	.725	.507	.225	.138	.670
8	7	.750	.430	.773	.700	.825	.596	.325	.088	.863
10	7	.770	.447	.809	.752	.735	.543	.235	.133	.623
10	8	.715	.430	.761	.677	.765	.548	.265	.118	.744
10	9	.770	.396	.805	.769	.830	.581	.330	.085	.877
15	11	.855	.701	.803	.744	.750	.555	.250	.125	.750
15	12	.725	.435	.800	.739	.820	.630	.320	.090	.845
15	13	.795	.551	.791	.709	.840	.646	.320	.090	.858
20	16	.745	.470	.813	.768	.850	.688	.350	.075	.891
20	18	.860	.526	.861	.846	.835	.558	.335	.083	.907
40	32	.840	.659	.842	.821	.880	.740	.380	.060	.939

Table 13  
 All Simulated Test Lengths and Associated Cut-Offs For Simulation Nine:  
 Equal Weights

No. of Items	Reliability			Validity						
	Cut-Off	Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency
2	1	.890	.361	.891	.889	.545	.251	.045	.288	-.116
2	2	.700	.372	.882	.577	.715	.458	.215	.143	.582
4	3	.655	.086	.791	.673	.665	.387	.165	.168	.413
4	4	.645	.279	.851	.601	.735	.467	.235	.133	.665
6	4	.800	.308	.835	.804	.775	.557	.275	.113	.561
6	5	.730	.429	.771	.634	.740	.499	.240	.130	.653
8	6	.790	.378	.811	.765	.750	.444	.250	.125	.607
8	7	.700	.403	.759	.654	.720	.438	.220	.140	.694
10	7	.810	.472	.809	.775	.750	.551	.250	.125	.548
10	8	.740	.437	.789	.711	.800	.587	.300	.100	.753
10	9	.705	.383	.770	.693	.785	.560	.285	.108	.818
15	11	.815	.566	.825	.784	.795	.632	.295	.103	.768
15	12	.785	.551	.795	.755	.810	.628	.310	.095	.792
15	13	.730	.456	.772	.717	.800	.608	.300	.100	.846
20	16	.805	.602	.810	.742	.800	.599	.300	.100	.821
20	18	.770	.501	.798	.743	.765	.567	.265	.118	.760
40	32	.815	.610	.837	.808	.855	.715	.355	.073	.884

Table 14  
 All Simulated Test Lengths and Associated Cut-Offs For Simulation Ten:  
 Equal Weights

No. of Items	Cut-Off	Reliability				Validity				
		Prop. Agree	Kappa	Subkoviak (1)	Subkoviak (2)	Prop. Agree	Validity	Utility	Disutility	Unweighted Efficiency
2	1	.720	.090	.841	.729	.380	.210	-.120	.310	-.180
2	2	.705	.230	.837	.713	.805	.193	.305	.098	.782
4	3	.680	.312	.735	.584	.650	.227	.150	.175	.536
4	4	.835	.143	.853	.841	.855	.159	.355	.073	.891
6	4	.740	.440	.757	.669	.690	.265	.190	.155	.651
6	5	.865	.413	.863	.858	.900	.482	.400	.050	.900
8	6	.805	.532	.801	.737	.735	.353	.235	.133	.771
8	7	.820	.449	.811	.801	.880	.572	.380	.060	.933
10	7	.730	.298	.811	.770	.775	.325	.275	.113	.794
10	8	.860	.466	.872	.866	.900	.546	.400	.050	.929
10	9	.890	.029	.894	.929	.945	.393	.445	.028	.974
15	11	.805	.578	.825	.787	.805	.597	.305	.098	.814
15	12	.920	.294	.892	.913	.940	.441	.440	.030	.970
15	13	.950	.519	.940	.954	.960	.582	.460	.020	.986
20	16	.930	.475	.902	.887	.945	.450	.445	.027	.975
20	18	.970	.556	.967	.978	.955	.456	.455	.023	.993
40	32	.945	.672	.929	.937	.955	.713	.455	.023	.985

Appendix Two

## Insurance Licensing Study

Figure 1: Letter to Panel Members

Figure 2: Overview of Tasks

Figure 3: Question Rating Form

Figure 4: Content Rating Form - Life

Figure 5: Content Rating Form - Accident and Health

Figure 6: Content Rating Form - Property

Figure 7: Content Rating Form - Causalty

Figure 1

EDUCATIONAL TESTING SERVICE



PRINCETON, N.J. 08541

009-021-0000

CABLE-EDUCTESTSVC

CENTER FOR OCCUPATIONAL  
AND PROFESSIONAL ASSESSMENTMemorandum for: INSURANCE STUDY  
PANEL MEMBERSSubject: Some Sample Documents  
for your Review

Date: October 6, 1978

We are delighted you will be assisting us in our very important study to evaluate the content of the Part 1 insurance licensing tests. Enclosed for your general information are samples of the documents you will be using in performing two tasks:

1. To examine the description of the test content (Content Outline) in relation to the test questions and to ascertain whether or not these content areas and questions are appropriate for a minimal competency test.
2. To examine individual test questions and to make judgements about the success of minimally competent persons on each test question.

The information from the panels will be used to arrive at a statistical estimate of the scores that a minimally knowledgeable individual for licensure in each line of insurance might expect to achieve.

Enclosed are the following:

1. Overview of Tasks
2. Question-Writing Guide
3. Sample Life Content Rating Form and Instructions
4. Sample Property Content Rating Form and Instructions
5. Sample Question Rating Form

We greatly look forward to seeing you on October 17, 1978.

Enclosures

October, 1978

Figure 2  
INSURANCE LICENSING STUDY

Overview of Tasks

The study in which you have been asked to participate will involve the collecting and analyzing of judgemental data to identify and validate the pass/fail decision in the Insurance Licensing Testing program. The results of this study in addition to the statistical analysis performed on each question and the test as a whole after it is given to large groups of people will be of assistance to the Commissioners of Insurance in the execution of their statutory responsibility to determine the minimum competence of individuals wishing to be licensed. Your judgements will be combined with judgements made by other insurance professionals to derive an estimate of the probable test performance of individuals wishing to be licensed as insurance agents.

As you know, a meeting will be held at the Holiday Inn/O'Hare Kennedy on October 17, 1978. The judgements, however, will be made individually and independently; members of the same panel will not confer as a group, nor will any member be informed of the judgements made by any other individual member. The judgements of all members of a panel will be combined statistically by ETS to arrive at a summary judgement for the panel about each question. The summary results for the questions also will be combined, and the final summary results will be published in a report describing the study and its findings or conclusions.

Several of the items in this mailing are intended to help you prepare for your tasks:

1. Content Outlines of each of the tests you will be reviewing were mailed to you previously. These outlines provide a blueprint of the major topics included in the tests and indicates the relative emphasis or number of questions that are given to each. They will serve to familiarize you with the general content of the test before you see the test questions themselves.

OVER  
↑

- 2 -

2. Question-Writing Guide. This abbreviated booklet describes the basic question formats used in the tests and general guidelines for preparing questions that are concise, unambiguous, and clearly stated. It is hoped that your review of the Guide will be of assistance to you in estimating the appropriateness of the actual test questions you will be seeing.
3. Rating Form. At our meeting in Chicago, you will be given a set of test questions and asked to make judgements on these forms. Please study the instructions and the form carefully before our meeting so that you can ask any questions about the tasks during the orientation session at the start of the meeting.

Before coming to the meeting, please give some thought to the kinds of abilities and knowledges that are essential to a person demonstrating minimum competency in insurance. You might think of particular situations where those abilities and knowledges are demonstrated for the protection of the public welfare. We will discuss this concern early in our meeting on October 17, 1978.

October, 1978

Figure 3

## INSURANCE LICENSING STUDY

## Question Rating Form

Name of Panel Member \_\_\_\_\_

Test \_\_\_\_\_

Question Number	Estimated Percentage of Minimally Knowledgeable Individuals Who Know Answer to Question								Question Number	Estimated Percentage of Minimally Knowledgeable Individuals Who Know Answer to Question							
	5	20	40	60	75	90	95	DNK		5	20	40	60	75	90	95	DNK
1.	5	20	40	60	75	90	95	DNK	26.	5	20	40	60	75	90	95	DNK
2.	5	20	40	60	75	90	95	DNK	27.	5	20	40	60	75	90	95	DNK
3.	5	20	40	60	75	90	95	DNK	28.	5	20	40	60	75	90	95	DNK
4.	5	20	40	60	75	90	95	DNK	29.	5	20	40	60	75	90	95	DNK
5.	5	20	40	60	75	90	95	DNK	30.	5	20	40	60	75	90	95	DNK
6.	5	20	40	60	75	90	95	DNK	31.	5	20	40	60	75	90	95	DNK
7.	5	20	40	60	75	90	95	DNK	32.	5	20	40	60	75	90	95	DNK
8.	5	20	40	60	75	90	95	DNK	33.	5	20	40	60	75	90	95	DNK
9.	5	20	40	60	75	90	95	DNK	34.	5	20	40	60	75	90	95	DNK
10.	5	20	40	60	75	90	95	DNK	35.	5	20	40	60	75	90	95	DNK
11.	5	20	40	60	75	90	95	DNK	36.	5	20	40	60	75	90	95	DNK
12.	5	20	40	60	75	90	95	DNK	37.	5	20	40	60	75	90	95	DNK
13.	5	20	40	60	75	90	95	DNK	38.	5	20	40	60	75	90	95	DNK
14.	5	20	40	60	75	90	95	DNK	39.	5	20	40	60	75	90	95	DNK
15.	5	20	40	60	75	90	95	DNK	40.	5	20	40	60	75	90	95	DNK
16.	5	20	40	60	75	90	95	DNK	41.	5	20	40	60	75	90	95	DNK
17.	5	20	40	60	75	90	95	DNK	42.	5	20	40	60	75	90	95	DNK
18.	5	20	40	60	75	90	95	DNK	43.	5	20	40	60	75	90	95	DNK
19.	5	20	40	60	75	90	95	DNK	44.	5	20	40	60	75	90	95	DNK
20.	5	20	40	60	75	90	95	DNK	45.	5	20	40	60	75	90	95	DNK
21.	5	20	40	60	75	90	95	DNK	46.	5	20	40	60	75	90	95	DNK
22.	5	20	40	60	75	90	95	DNK	47.	5	20	40	60	75	90	95	DNK
23.	5	20	40	60	75	90	95	DNK	48.	5	20	40	60	75	90	95	DNK
24.	5	20	40	60	75	90	95	DNK	49.	5	20	40	60	75	90	95	DNK
25.	5	20	40	60	75	90	95	DNK	50.	5	20	40	60	75	90	95	DNK



October, 1978

Figure 4

INSURANCE LICENSING STUDY  
Content Rating Form  
Life Test

Name of Panel Member \_\_\_\_\_

Question Number	Is the question appropriate to the content area?			Question Number	Is the question appropriate to the content area?		
1.	YES	NO	DNK	26.	YES	NO	DNK
2.	YES	NO	DNK	27.	YES	NO	DNK
3.	YES	NO	DNK	28.	YES	NO	DNK
4.	YES	NO	DNK	29.	YES	NO	DNK
5.	YES	NO	DNK	30.	YES	NO	DNK
6.	YES	NO	DNK	31.	YES	NO	DNK
7.	YES	NO	DNK	32.	YES	NO	DNK
8.	YES	NO	DNK	33.	YES	NO	DNK
9.	YES	NO	DNK	34.	YES	NO	DNK
10.	YES	NO	DNK	35.	YES	NO	DNK
11.	YES	NO	DNK	36.	YES	NO	DNK
12.	YES	NO	DNK	37.	YES	NO	DNK
13.	YES	NO	DNK	38.	YES	NO	DNK
14.	YES	NO	DNK	39.	YES	NO	DNK
15.	YES	NO	DNK	40.	YES	NO	DNK
16.	YES	NO	DNK	41.	YES	NO	DNK
17.	YES	NO	DNK	42.	YES	NO	DNK
18.	YES	NO	DNK	43.	YES	NO	DNK
19.	YES	NO	DNK	44.	YES	NO	DNK
20.	YES	NO	DNK	45.	YES	NO	DNK
21.	YES	NO	DNK	46.	YES	NO	DNK
22.	YES	NO	DNK	47.	YES	NO	DNK
23.	YES	NO	DNK	48.	YES	NO	DNK
24.	YES	NO	DNK	49.	YES	NO	DNK
25.	YES	NO	DNK	50.	YES	NO	DNK

October, 1978

Figure 5

INSURANCE LICENSING STUDY  
 Content Rating Form  
Accident and Health Test

Name of Panel Member \_\_\_\_\_

<u>Question Number</u>	<u>Is the question appropriate to the content area?</u>			<u>Question Number</u>	<u>Is the question appropriate to the content area?</u>		
1.	YES	NO	DNK	26.	YES	NO	DNK
2.	YES	NO	DNK	27.	YES	NO	DNK
3.	YES	NO	DNK	28.	YES	NO	DNK
4.	YES	NO	DNK	29.	YES	NO	DNK
5.	YES	NO	DNK	30.	YES	NO	DNK
6.	YES	NO	DNK	31.	YES	NO	DNK
7.	YES	NO	DNK	32.	YES	NO	DNK
8.	YES	NO	DNK	33.	YES	NO	DNK
9.	YES	NO	DNK	34.	YES	NO	DNK
10.	YES	NO	DNK	35.	YES	NO	DNK
11.	YES	NO	DNK	36.	YES	NO	DNK
12.	YES	NO	DNK	37.	YES	NO	DNK
13.	YES	NO	DNK	38.	YES	NO	DNK
14.	YES	NO	DNK	39.	YES	NO	DNK
15.	YES	NO	DNK	40.	YES	NO	DNK
16.	YES	NO	DNK	41.	YES	NO	DNK
17.	YES	NO	DNK	42.	YES	NO	DNK
18.	YES	NO	DNK	43.	YES	NO	DNK
19.	YES	NO	DNK	44.	YES	NO	DNK
20.	YES	NO	DNK	45.	YES	NO	DNK
21.	YES	NO	DNK	46.	YES	NO	DNK
22.	YES	NO	DNK	47.	YES	NO	DNK
23.	YES	NO	DNK	48.	YES	NO	DNK
24.	YES	NO	DNK	49.	YES	NO	DNK
25.	YES	NO	DNK	50.	YES	NO	DNK

October, 1978

Figure 6

INSURANCE LICENSING STUDY  
Content Rating Form  
Property Test

Name of Panel Member \_\_\_\_\_

Question Number	Is the question appropriate to the content area?			Question Number	Is the question appropriate to the content area?		
1.	YES	NO	DNK	26.	YES	NO	DNK
2.	YES	NO	DNK	27.	YES	NO	DNK
3.	YES	NO	DNK	28.	YES	NO	DNK
4.	YES	NO	DNK	29.	YES	NO	DNK
5.	YES	NO	DNK	30.	YES	NO	DNK
6.	YES	NO	DNK	31.	YES	NO	DNK
7.	YES	NO	DNK	32.	YES	NO	DNK
8.	YES	NO	DNK	33.	YES	NO	DNK
9.	YES	NO	DNK	34.	YES	NO	DNK
10.	YES	NO	DNK	35.	YES	NO	DNK
11.	YES	NO	DNK	36.	YES	NO	DNK
12.	YES	NO	DNK	37.	YES	NO	DNK
13.	YES	NO	DNK	38.	YES	NO	DNK
14.	YES	NO	DNK	39.	YES	NO	DNK
15.	YES	NO	DNK	40.	YES	NO	DNK
16.	YES	NO	DNK	41.	YES	NO	DNK
17.	YES	NO	DNK	42.	YES	NO	DNK
18.	YES	NO	DNK	43.	YES	NO	DNK
19.	YES	NO	DNK	44.	YES	NO	DNK
20.	YES	NO	DNK	45.	YES	NO	DNK
21.	YES	NO	DNK	46.	YES	NO	DNK
22.	YES	NO	DNK	47.	YES	NO	DNK
23.	YES	NO	DNK	48.	YES	NO	DNK
24.	YES	NO	DNK	49.	YES	NO	DNK
25.	YES	NO	DNK	50.	YES	NO	DNK

-2-  
Property Test

Questions 51-60 refer to the major topics covered in the Property Test.

51. Seventeen questions (#1-17) deal with Basic Principles and Concepts in Property insurance. (See pages 4-5 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES                      NO                      DNK
52. Twelve questions (#18-29) deal with the Standard Fire Policy. (See page 5 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES                      NO                      DNK
53. Ten questions (#30-39) deal with Forms and Endorsements attached to the Standard Fire Policy. (See page 5 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES                      NO                      DNK
54. Ten questions (#40-49) deal with Package Policies. (See pages 5-6 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES                      NO                      DNK
55. One question (#50) deals with the Nature and Purpose of National Flood Insurance. (See page 6 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES                      NO                      DNK
56. Of the seventeen questions covering Basic Principles, how many of these questions do you think a minimally competent person will answer correctly?
- \_\_\_\_\_
57. Of the twelve questions covering the Standard Fire Policy, how many of these questions do you think a minimally competent person will answer correctly?
- \_\_\_\_\_
58. Of the ten questions covering Forms and Endorsements, how many of these questions do you think a minimally competent person will answer correctly?
- \_\_\_\_\_

-3-

Property Test

59. Of the ten questions covering Package Policies, how many of these questions do you think a minimally competent person will answer correctly?

---

60. Would the minimally competent person answer the one question on National Flood Insurance correctly?

 Yes No

October, 1978

Figure 7  
INSURANCE LICENSING STUDY  
Content Rating Form  
Casualty Test

Name of Panel Member \_\_\_\_\_

Question Number	Is the question appropriate to the content area?			Question Number	Is the question appropriate to the content area?		
1.	YES	NO	DNK	26.	YES	NO	DNK
2.	YES	NO	DNK	27.	YES	NO	DNK
3.	YES	NO	DNK	28.	YES	NO	DNK
4.	YES	NO	DNK	29.	YES	NO	DNK
5.	YES	NO	DNK	30.	YES	NO	DNK
6.	YES	NO	DNK	31.	YES	NO	DNK
7.	YES	NO	DNK	32.	YES	NO	DNK
8.	YES	NO	DNK	33.	YES	NO	DNK
9.	YES	NO	DNK	34.	YES	NO	DNK
10.	YES	NO	DNK	35.	YES	NO	DNK
11.	YES	NO	DNK	36.	YES	NO	DNK
12.	YES	NO	DNK	37.	YES	NO	DNK
13.	YES	NO	DNK	38.	YES	NO	DNK
14.	YES	NO	DNK	39.	YES	NO	DNK
15.	YES	NO	DNK	40.	YES	NO	DNK
16.	YES	NO	DNK	41.	YES	NO	DNK
17.	YES	NO	DNK	42.	YES	NO	DNK
18.	YES	NO	DNK	43.	YES	NO	DNK
19.	YES	NO	DNK	44.	YES	NO	DNK
20.	YES	NO	DNK	45.	YES	NO	DNK
21.	YES	NO	DNK	46.	YES	NO	DNK
22.	YES	NO	DNK	47.	YES	NO	DNK
23.	YES	NO	DNK	48.	YES	NO	DNK
24.	YES	NO	DNK	49.	YES	NO	DNK
25.	YES	NO	DNK	50.	YES	NO	DNK

-2-

Casualty Test

Questions 51-60 refer to the major topics covered in the Casualty Test.

51. Fifteen questions (#1-15) deal with Basic Principles and Concepts in Casualty insurance. (See page 6 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES NO DNK
52. Fifteen questions (#16-30) deal with Basic Concepts of Auto Insurance. (See page 6 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES NO DNK
53. Thirteen questions (#31-43) deal with General Liability Contracts. (See pages 6-7 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES NO DNK
54. Five questions (#44-48) deal with basic concepts of Crime Insurance. (See page 7 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES NO DNK
55. Two questions (#49-50) deal with General Principles of Suretyship. (See page 7 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES NO DNK
56. Of the fifteen questions covering Basics, how many of these questions do you think a minimally competent person will answer correctly?
- 
57. Of the fifteen questions covering Basics of Auto Insurance, how many of these questions do you think a minimally competent person will answer correctly?
- 
58. Of the thirteen questions covering General Liability Contracts, how many of these questions do you think a minimally competent person will answer correctly?
-

-3-

Casualty Test

59. Of the five questions covering Crime Insurance, how many of these questions do you think a minimally competent person will answer correctly?

---

60. Of the three questions covering Principles of Suretyship, how many of these questions do you think a minimally competent person will answer correctly?

---



-2-

Accident and Health Test

Questions 51-58 refer to the major topics covered in the Accident and Health Test.

51. Ten questions (#1-10) deal with Basics in Accident and Health insurance. (See page 9 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES            NO            DNK
52. Twenty questions (#11-30) deal with Individual Accident and Health Provisions. (See pages 9-10 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES            NO            DNK
53. Fifteen questions (#31-45) deal with Types of Coverage. (See page 10 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES            NO            DNK
54. Five questions (#46-50) deal with Types of Contracts. (See page 10 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES            NO            DNK
55. Of the ten questions covering Basics in Accident and Health insurance, how many of these questions do you think a minimally competent person will answer correctly?
- \_\_\_\_\_
56. Of the twenty questions covering Individual Accident and Health Provisions, how many of these questions do you think a minimally competent person will answer correctly?
- \_\_\_\_\_
57. Of the fifteen questions dealing with Types of Coverage, how many of these questions do you think a minimally competent person will answer correctly?
- \_\_\_\_\_
58. Of the five questions dealing with Types of Contracts, how many of these questions do you think a minimally competent person will answer correctly?
- \_\_\_\_\_

-2-  
Life Test

Questions 51-56 refer to the major topics covered in the Life Test.

51. Ten questions (#1-10) deal with Basics in Life insurance. (See page 8 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES                      NO                      DNK
52. Twenty questions (#11-30) deal with Life Insurance Provisions. (See pages 8-9 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES                      NO                      DNK
53. Twenty questions (#31-50) deal with Kinds of Life Insurance and Annuities. (See page 9 of the Bulletin for a full breakdown of this section.) Is this content area appropriate for a minimum competency test?
- YES                      NO                      DNK
54. Of the ten questions covering Basics in Life insurance, how many of these questions do you think a minimally competent person will answer correctly?
- \_\_\_\_\_
55. Of the twenty questions covering Life Insurance Provisions, how many of these questions do you think a minimally competent person will answer correctly?
- \_\_\_\_\_
56. Of the twenty questions covering Kinds of Life Insurance and Annuities, how many of these questions do you think a minimally competent person will answer correctly?
- \_\_\_\_\_

