DOCUMENT RESUME

ED 092 593 95 TM 003 705

AUTHOR	Brennan, Robert L.
TITLE	The Evaluation of Mastery Test Items. Final Report.
SPONS AGENCY	National Center for Educational Research and Development (DHEW/OE), Washington, D.C.
BUREAU NO	BR-2-B-118
PUB DATE	Jan 74
GRANT	OEG-2-2-2B118
NOTE	255p.
EDRS PRICE	MF-$0.75 HC-$12.60 PLUS POSTAGE
DESCRIPTORS	Computer Programs; Confidence Testing; *Criterion Referenced Tests; *Item Analysis; *Literature Reviews; *Scoring; Scoring Formulas; Standard Error of Measurement; Test Construction; Testing; *Test Reliability; Test Validity; True Scores
IDENTIFIERS	Classical Test Theory; *Mastery Testing; Test Items

ABSTRACT

The first four chapters of this report primarily provide an extensive, critical review of the literature with regard to selected aspects of the criterion-referenced and mastery testing fields. Major topics treated include: (a) definitions, distinctions, and background, (b) the relevance of classical test theory, (c) validity and procedures for test construction, and (d) test reliability. Chapter V provides a treatment of criterion-referenced and mastery item analysis and revision procedures when items are scored in the classical correct/wrong manner. Chapter VI treats an alternative to the classical procedure for administering and scoring items. This procedure employs the subjective probabilities typically associated with confidence testing in order to obtain pseudo-classical scores. These scores, which have not been considered elsewhere, appear to be very useful for item analysis purposes in that they have most of the advantages and few of the disadvantages of both classical scores and subjective probabilities. Chapter VII provides an analysis of a set of data collected to illustrate many of the statistics and procedures discussed in Chapter V and VI. One of the appendices provides the manual an extensive test scoring and item analysis program that uses student subjective probabilities as input. (Author)

FINAL REPORT

Project No. 2B118
Grant No. OEG-2-2-2B118

Robert L. Brennan
Department of Education
SUNY at Stony Brook
Stony Brook, N.Y. 11790

THE EVALUATION OF MASTERY TEST ITEMS

January, 1974

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Office of Education
National Center for Educational Research and Development
(Regional Research Program)

Abstract

The first four chapters of this report primarily provide an extensive, critical review of the literature with regard to selected aspects of the criterion-referenced and mastery testing fields. Major topics treated include: (a) definitions, distinctions, and background, (b) the relevance of classical test theory, (c) validity and procedures for test construction, and (d) test reliability.

Chapter V provides a treatment of criterion-referenced and mastery item analysis and revision procedures when items are scored in the classical correct/wrong manner. Chapter VI treats an alternative to the classical procedure for administering and scoring items. This procedure employs the subjective probabilities typically associated with confidence testing in order to obtain pseudo-classical scores. These scores, which have not been considered elsewhere, appear to be very useful for item analysis purposes in that they have most of the advantages and few of the disadvantages of both classical scores and subjective probabilities.

Chapter VII provides an analysis of a set of data collected to illustrate many of the statistics and procedures discussed in Chapters V and VI, especially.

One of the appendices to this report provides the manual for an extensive test scoring and item analysis program that uses student subjective probabilities as input.

Final Report

Project No. 2B118
Grant No. OEG-2-2-2B118

THE EVALUATION OF MASTERY TEST ITEMS

Robert L. Brennan

State University of New York
at Stony Brook

Stony Brook, New York

January, 1974

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
National Center for Educational Research and Development

Preface

The author would like to acknowledge assistance received by him from several people. Ms. Sandra Radoff and Ms. Gail Ironson were very helpful in collecting references and performing other supportive tasks; Dr. Michael Kane offered several comments and critical observations that considerably influenced the development of Chapter VI; and Mr. Joe Crick proofread and commented upon several chapters contained in this report. I would also like to express my gratitude to Dr. Emir Shuford for fostering my interest in admissible probability measurement -- a topic than influenced much of the work reported here. Finally, I am in debt to Dr. David McMullen for his criticism of several aspects of this manuscript.

Table of Contents

List of Tables

Number Title Page

List of Figures

CHAPTER I

Introduction

Rationale for and Overview of Research Reported Here

In the last decade there has been considerable discussic.), debate, research, and development surrounding criterion-referenced testing and mastery testing. Inevitably, the issues have been discussed from philosophical, theoretical, and practical points of view. Some persons have been primarily concerned about comparing these new testing techniques with norm-referenced techniques; some have argued that there are no important differences among these techniques; others have argued that there are important differences; and still others have assumed that there are important differences and proceeded from there.

Thus, the modus operandi among researchers who have worked in the areas of criterion-referenced and mastery testing has differed considerably, and this is probably desirable, in general. However, this fact, the relative youth of these testing techniques, their apparent popularity, and their somewhat unrestrained use have all interacted to confuse certain issues and to render very difficult an answer to the question, "What do we know about criterion-referenced testing?" Relatively little of what we know is found in textbooks or even in the popular journals that treat measurement and testing. As might be expected, some of the best work is found in unpublished manuscripts and reports.

Thus, one purpose of this document is to provide an overview of the literature on criterion-referenced and mastery testing, especially with regard to statistical measures, criteria, and procedures for criterion-referenced reliability, validity, and item analysis. Any such review of the literature is bound to be somewhat biased by the author's subjective judgment, and it is virtually impossible to reference all the work performed in any of these areas. However, an effort has been made to identify the most important, or potentially important references.

Another purpose of this report is to discuss procedures for identifying criterion-referenced and mastery test items that requ.re revision. It is the feeling of this author that this is a fundamentally important topic, in that: (a)consideration of the problems involved here helps to clarify the important issues in criterion-referenced reliability and validity, and (b) any measurement technique can only be valid and useful to the extent that the measuremetn instrument, and, hence, the test items are at least minimally acceptable.

The proposal that formed the initial statement of the research reported here also indicated that several different kinds of item administration and scoring procedures would be considered in terms of their applicability for criterion-referenced and mastery testing. At the time the proposal was writtern, this author felt that the typical correct/wrong scoring procedure for objective items left much to be desired, especially for many applications of criterion-referenced testing. In particular, this author felt that confidence testing, or one of its variants, might offer significant advantages to criterion-referenced testing. The research discussed in the later chapters of this report seems to support these beliefs.

Definitions, Distinctions, and Background

Norm-referenced testing. Measurement theory traditionally has been concerned with the accurate estimation and interpretation of an individual's score in relation to the scores of other individuals who have taken, or who might potentially take, a given test. In fact, many psychometricians have historically taken the position that "a test (that is) not discriminating among examinees ... is not a useful measuring instrument (Lord and Novick, 1968, p. 252)." However, it should be noted that very few psychometricians define measurement in a manner that necessitates this discriminating function of a test. (See, for example, the definition of measurement provided by Lord and Novick, 1968, p. 17.) In other words, historically, most psychometricians have concerned themselves with tests whose purpose is to maximally discriminate among subjects with regard to some underlying characteristic, trait, or construct; hence, the ability of a test to provide a basis for making statistical statements about the distinctions among students has, for many, become an operational definition of "useful measurement instrument." Furthermore, this point of view has necessitated that the interpretation of a student's score be "... dependent on the relative position of the score in comparison with other scores (Popham and Husek, 1969, p. 3)." Tests of this kind are currently referred to as norm-referenced tests. The term "norm-referenced" is somewhat inappropriate in that the "norm group" in traditional test theory usually has a specific definition or connotation that is not necessarily consistent with the term "norm-referenced"; however, here, as elsewhare in this report, our concern is with describing terms in an unambiguous manner, not changing their names.

Criterion-referenced testing -- background. Now, there can be no argument about the usefulness of norm-referenced testing; however, researchers for a number of years have noted that certain purposes and uses of tests do not fit very well into a norm-referenced framework. Flanagan (1951) and Gardner (1962) pointed out some distinctions between what are now called norm-referenced and criterion-referenced tests; Ebel's (1962) work on "content standard scores" is also frequently referenced as a precursor to criterion-referenced testing. However, Glaser (1963) and Glaser and Klaus (1962) are the earliest references that specifically consider criterion-referenced tests, as such; the latter, in the opinion of this author, is still one of the best introductions to the distinctions between norm-referenced and criterion-referenced tests.

It is interesting to note the historical proximity between criterion-referenced testing and programmed instruction, which provided a motivating factor in the development of new instructional systems and educational technology. This chronological proximity is probably not mere coincidence, since, as Coulson adn Cogswell (1965) note, changes in testing procedures are a natural consequence of changes in teaching method. In fact, most criterion-referenced testing is closely associated with some kind of instruction, especially individualized or adaptive instruction. (See, for example, Nitko, 1971, and Hambleton, 1973.)

There are several lessons to be learned from this frequently occuring relationship between criterion-referenced testing and instruction. First, it should be noted that this relationship can easily confound the interpretability of criterion-referenced measurements. In fact, one of the difficulties with most of the literature is a needless confusion of instruction and measurement. This is not to say that instruction and measurement cannot and should not interact. This author has even stated elsewhere that many of the problems in instruction will not be resolved until fundamental issues in measurement are adequately treated (Brennan, 1973b). However, at least at the present time, in the opinion of this author, it is necessary to recognize the distinctions between measurement and instruction, if we are to advance the cause of either. The relationship between criterion-referenced and instruction may also provide an explanation for the rather uneven interest, if not the apathy, of many psychometricians with regard to criterion-referenced testing. From a practical point of view, good criterion-referenced test data for a reasonably large number of subjects is quite rare, or not readily available to psychometricians for analysis. For one thing, the collection and analysis of criterion-referenced test data is often somewhat over-shadowed by the day-to-day exigencies of providing instruction to students. Also, many psychometricians work in environments that remove them from the testing issues that often arise in instructional contexts; hence, such psychometricians are often removed from the issues that motivate much of the work in criterion-referenced testing.

Criterion-referenced testing -- definitions. Many definitions of a criterion-referenced test have been proposed in the literature. For example:

> "A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed (Harris and Stewart, 1971, p. 1)."
>
> "A criterion-referenced test is one composed of items keyed to a set of behavioral o^iectives (Ivens, 1970, p. 2)."
>
> "A criterion-referenced test is one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards (Glaser and Nitko, 1971)."

The last definition seems to be the one that is most universally accepted, the second is one of the most general definitions, and the first is one of the most specific. This author prefers the last definition; however, many criterion-referenced tests appear to satisfy all definitions; and, therefore, arguments about the "best" definition may be of more theoretical than practical concern. From another point of view, however, it should be noted that some tests which are criterion-referenced under one definition may not be criterion-referenced under another definition.

Criterion-referenced testing -- characteristics. Among the most frequently cited characteristics of a criterion-referenced test are: (a) test items are associated with specific behavioral objectives, (b) the resulting measurement scale is an absolute, as opposed to, a relative scale,(c) a student's score is capable of being interpreted independent of the scores of other subjects, and (d) there is a specified behavioral criterion (or criteria) for acceptable performance. It is worth considering some of these characteristics in more detail.

From a practical point of view, criterion-referenced test items are almost always claimed to be associated with specific objectives. However, this author does not believe that anything in Glaser and Nitko's (1971) definition of a criterion-referenced test necessitates that criterion-referenced items must necessarily be associated with the typical kinds of presently available, explicitly stated behavioral objectives. This is not an argument against behavioral objectives; rather it is an admonition not to needlessly constrain the definition of a criterion-referenced test by demanding that criterion-referenced test items reflect particular kinds of behavioral objectives. Nevertheless, it is critical that some behavioral criterion for acceptable performance be specified.

Glaser (1963) discusses the issue of absolute versus relative standards in the following terms:

> "The scores obtained from an achievement test provide primarily two kinds of information. One is the degree to which the student has attained criterion performance, for example, whether he can satisfactorily prepare an experimental report, or solve certain kinds of work problems in arithmetic. The second kind of information that an achievement test score provides is the relative ordering of individuals with respect to their test performance, for example, whether student A can solve his problems more quickly that student B. The principal difference between these two kinds of information lies in the standard used as a reference. What can be called criterion-referenced measures depend upon an absolute standard of quality, while what can be termed norm-referenced measures depend upon a relative standard (Glaser, 1963, p. 2)."

We stated above that norm-referenced tests are specifically constructed to yield scores that allow for maximum discrimination among subjects. More precisely, such measures are intended to provide a basis for making distinctions among subjects over a continuum of ability. Eventhough the interpretation of a subject's criterion-referenced score is independent of the scores obtained by other subjects; it is not quite true to say that all criterion-referenced tests are not intended to identify differences among subjects. However, the differences to be identified are of a very specific nature. That is, a criterion-referenced test is often intended to distinguish between two groups of subjects: those who have and those who have not achieved the specified performance standard. For the most part, criterion-referenced tests that have this intended purpose fall into the realm of mastery testing.

Thus, norm-referenced and criterion-referenced tests differ with regard to the desired nature of the discriminations among subjects. An analogy may help clarify this point. In describing the length of a table, I may say that it is either greater than or not greater than six feet long; or I may say that it is longer than 80 percent of the tables in the school cafeteria. The latter is analogous to the norm-referenced kind of discrimination. Also, note that the criterion-referenced statement makes use of an absolute measurement scale, while the norm-referenced statement does not.

Mastery testing. That part of Glaser and Nitko's (1971) definition of a criterion-referenced test that refers to a "specified performance standard" has been a subject of considerable confusion and misunderstanding in the literature. The standard should be specified and it should be amenable to measurement of some kind (hence, the word "performance"); but the standard need not be a single score, the standard need not be high, and certainly the standard need n t be perfect mastery, or anything close to perfect mastery. Now, it is often true that the standard chosen is a single "high" score, and, thus, in many cases, there is little operational difference between a criterion-referenced test and a mastery test; however, the difference between these two kinds of tests is a potentially real and important one. This distinction should be recognized even if, in particular circumstances, the distinction is not made. For example, the tests used in the National Assessment Program (see Merwin and Womer, 1969) can be considered to be criterion-referenced, but they are not a typical example of mastery tests. In this report we will make the distinction between criterion-referenced and mastery testing when we deem it to be critical; otherwise, we will use the term "criterion-referenced" instead of "mastery," since the latter is a special case of the former.

The impetus for, and original work in, mastery testing was presented by Bloom (1968) as part of a general model for mastery learning. Perhaps the best-known references on the topic are Bloom (1971) and Block (1971). The latter presents a review of the literature which has recently been updated by the same author (Block, 1973). From a measurement viewpoint the outstanding issue in a discussion of mastery testing is the cut-off value (cutting score, passing score, mastery cut-off, or criterion) chosen as the basis for classifying stuents as masters or non-masters. Emrick (1971), Kriewall (1969), and Millman (1972) have all treated this issue to some extent. However, there seems to be a subtle difference between Glaser and Nitko's "specified performance standard" and the basis upon which some persons recommend choosing a mastery cutting score. At least sometimes, the mastery cutting scores appear to be based partially upon characteristics of the test score distribution. Such procedures, in the opinion of this author, run the risk of confounding the definition of mastery with the irrelevant information provided by the test score distribution. For example, taken to extremes, such a procedure might, after the fact, classify all persons above the median as masters, in which case, mastery learning is guaranteed to be effective (and not effective) for fifty percent of the students.

The word "criterion." Another issue with regard to general terminology and background for this report concerns the word "criterion." This word, for some time, has had several denotations or connotations in test theory; and with the advent ofcriterion-referenced and mastery testing the potential ambiguities have increased. In classical test theory, the word "criterion" usually refers to some external measure that provides a standard against which a particular test is compared; in this sense, the word "criterion" is often associated with criterion, statistical, or empirical validity (Brown, 1970). Also, in both classical testing and mastery testing the word criterion is sometimes synonymous with a cutting score, cut-off score, or "acceptable" score magnitude. In criterion-referenced testing, the word "criterion" refers generically to "the standard (or criterion) against which a student's performance is compared (Glaser, 1963, p. 519)." Nitko (1971) discusses these distinctions in somewhat greater depth. Once these distinctions are recognized, the context of a given discussion usually resolves any ambiguities.

Criterion-referenced tests and scores. It is almost inconceivable that a z-score, T-score, stanine, or percentile rank would be a criterion-referenced score, whereas "number of items correct" or "proportion of items correct" *might* be. Nevertheless, the actual student score reported is, of itself, never sufficient to warrant saying that the score is criterion-referenced. Such a statement can be made only if the test is (or can be interpreted as) a criterion-referenced test *and* the score reported reflects the specified performance standard.

Merely viewing a test is not sufficient to identify it as criterion-referenced or norm-referenced; one must also know the manner in which it was constructed, the purpose for which it will be used, and the way in which student scores are constructed and interpreted. Furthermore, practically any test has the potential for being either criterion-referenced or norm-referenced. Thus, for example Ebel (1962) has suggested a procedure for deriving criterion-referenced information from a norm-referenced test.

Ebel's limitations.of criterion-referenced testing. Surprisingly, Ebel is also a generally vocal critic of criterion-referenced testing. In Ebel's view, the major limitations of criterion-referenced tests are: "(1) they do not tell us all we need to know about achievement, (2) they are difficult to obtain on any sound basis, and (3) they are necessary for only a small fraction of important educational achievements (Ebel, 1970, p. 8)." The last objection is, I think very ambiguous in that Ebel does not define what he means by "important educational achievements." As a counterexample, in my experience, most teachers and those working with instructional systems, when asked to characterize a "good" test for their purposes, invariably list characteristics of criterion-referenced tests. Ebel's first "limitation" is, at best, misdirected in that very few researchers would want to argue that any testing technique is likely to be sufficient in providing us with "all we need to know about achievement (italics ours)." Ebel's second limitation is at least partially true, but it is not true that is prohibitively difficult to obtain good criterion-referenced measurements. Finally, in the opinion of this author, Ebel's three supposed limitations of criterion-referenced measurement are equally, if not more, appropriate comments about norm-referenced measurement. But even if one agrees that Ebel's statement of limitations is valid, this, in itself, is not a justification for eliminating the use or development of criterion-referenced testing, as some might claim. The issue is not which kind of testing is better, but rather, which kind of testing is appropriate, under what circumstances, and for what purpose.

A Model for the Use of Achievement Data and Time Data in an Instructional System

Since criterion-referenced and mastery testing are often closely associated with an instructional system, it is desirable to consider the role of these testing techniques in an instructional system; at the same time it is useful to to consider the potential role of norm-referenced testing in an instructional. In this section, we briefly consider these issues; the reader is referred to Brennan (1973a) for a more complete discussion.

Here we restrict ourselves to a consideration of achievement data and time data for evaluating the cognitive aspects of an instructional system. Given the current state-of-the-art, one might argue that achievement data and time data often provide the most useful and interpretable information with regard to decision-making in an instructional system; nevertheless, it should be noted that a complete evaluation of an instructional system necessitates the collection and use of other types of data, as well.

Objective-related modules. One reason that so much of the literature on evaluating particular instructional systems lacks generalizability to other instructional systems is that the unit of analysis for the purpose of collecting data and making decisions is apt to vary considerably from system to system; and, often enough, the unit of analysis varies even within the same system.

In some systems the unit of analysis is merely the amount of instruction that occurs in some specified time period; in other systems the unit of analysis corresponds with the instruction for some group of objectives which are taught together in some sequence for pedagogical reasons. In both of these cases, the unit of analysis corresponds with obvious physical characteristics of the system, and, therefore, the unit of analysis typically involves a number of different instructional objectives. However, the kinds of decisions that must be made in evaluating and revising an instructional system necessitate a consideration of all of the data and instruction relating to each separate objective, no matter when the data are collected or where the instruction occurs within the system.

In short, the basic unit of analysis in an instructional system should be the objective. In order to emphasize this fact and facilitate the collection and analysis of data for decision-making, it is theoretically and practically useful to view an instructional system as consisting of a discrete number of objective-related modules. As employed here, the phrase "objective-related module" refers to all of those factors in an instructional system that are directly related to a particular instructional objective. Note especially that the term "module" is not used here as a descriptive characteristic of the physical layout of an instructional system. The central aspects of an objective-related module are the objective itself and the instruction intended to teach the objective. In addition, an objective-related module contains all of the data directly relevant to the particular objective.

This conception of an instructional system in terms of objective-related modules may appear too theoretical or too trivial, at first glance; however, for purposes of evaluation, the concept of an objective-related module has several advantages over many other ways to outline and describe an instructional system. First, and most importantly, this concept directly implies that the objective is the basic unit of analysis in an instructional system. Second, the objective-related module concept emphasizes the relationship between the objective, instruction, and data. Third, any instructional system can be described in terms of objective-related modules, regardless of how the instruction is sequenced or packaged. Fourth, the objective-related module concept greatly facilitates an understanding of many of the issues and problems surrounding the collection and use of data in instructional systems.

Purposes of data collection. Any discussion of data immediately raises two questions: for what purpose should such data be collected and what kind of data should be collected? Here we restrict the scope of these two questions to the domain of evaluating cognitive achievement in an instructional system.

In general, of course, one can say that data is collected in the environment of an instructional system for the purpose of evaluation, where, according to Stufflebeam (1971) "evaluation is the process of delineating, obtaining, and providing useful information for judging decision alternatives (p. 267)."

More specifically, one could say that data should be collected for the purposes of diagnostic, formative, and summative evaluation (Bloom, Hastings, and Madaus, 1971). If one considers evaluation as a decision-making process, then the diagnostic-formative-summative trichotomy refers primarily to potential decision-making functions of evaluation. However, we prefer to emphasize that the purpose of collecting data in the environment of an instructional system is to make decisions with regard to specific aspects of the instructional system, namely: (a) instruction, (b) students, and (c) test items. That is, we prefer to emphasize the object of the decision-making process, as opposed to its function. Emphasizing the object of the decision-making process seems to identify more clearly the specific nature of the decisions that typically need to be made in an on-going instructional system.

Decisions about instruction are usually of primary importance; i.e., one wants to assess the effects of instruction especially for the purpose of identifying instruction that requires revision. Such decisions are often viewed as part of the process of formative evaluation. In order to make decisions concerning whether or not instruction should be revised, we argue here that data should be obtained which can be used to determine instructional effectiveness, efficiency, and retention.

Decisions about students typically include decisions concerning student placement and certification. Such decisions are often viewed as part of the processes of diagnostic and summative evaluation, respectively.

Decisions about test items also need to be made in instructional systems. Specifically, one needs to determine the reliability and validity of tests used as part of the instructional system.

Types of data. One can identify at least eight different types of data for an objective-related module that provide meaningful sources of information for decision-making.

These types of data listed in the order in which they would usually be obtained, are as follows:

(a) Prerequisite test data, which indicates whether or not a student has the background characteristics (attainment of previous objectives, aptitude, etc.) thought to be necessary in order to achieve the objective for the module;

(b) Pretest data, which measures a student's performance on the objective prior to instruction;

(c) Instructional time, which is the length of time a student spends undergoing instruction for the objective;

(d) Criterion-referenced posttest data, which measures a student's performance on the objective immediately after instruction;

(e) Norm-referenced posttest data, which is collected immediately after instruction and measures student performance relative to the performance of other similar students;

(f) Retention time, which is the length of time intervening between the posttest (usually criterion-referenced) and a subsequent retention test (usually criterion-referenced;

(g) Criterion-referenced retention test data, which is collected some time after instruction and measures student performance on the objective for the module; and

(h) Norm-referenced retention test data, which is collected some time after instruction and measures student performance relative to the performance of other similar students.

It is often assumed that only criterion-referenced or mastery test data provide meaningful information for evaluation decisions with regard to instructional systems. Certainly criterion-referenced data is more important that norm-referenced data in the context of an instructional system; however, norm-referenced data sometimes provides useful additional information for decision-making (see Brennan, 1973a, for more detail concerning this issue).

A table for relating data type and use. These data for an objective-related module are displayed in Table 1.1 which, in addition, indicates those types of data that are of primary importance for making decisions with regard to instruction, students, and test items. In essence, Table 1.1 provides a kind of taxonomy of achievement and time data that are useful in evaluating instructional systems. It is of course quite possible that a particular

TABLE 1-1

A Taxonomy of Achievement Data and Time Data for Decision-Making in an Instructional System

Data used to make decisions with regard to:	Data from an objective-related module[1]							
	Pre-req. test	Pre-test	Inst. time	C.R. post-test	N.R. post-test	Ret. time	C.R. ret. test	N.R. ret. test
Instruction:								
Short-term effectiveness		*		*				
Long-term effectiveness		*					*	
Efficiency		*	*	*				
Retention				*		*	*	
Students:								
Placement .	*	*		*			*	
Certification								
Mastery		*		*			*	
Grading				*	*		*	*
Tests:								
Validity					*			*
Reliability	*	*		*	*		*	*

[1] The data are listed from left to right in the chronological order that they would usually be collected. Asterisks indicate the principal kinds of data appropriatę for particular decicions.

objective-related module may not contain all of the data indicated in Table 1-1. It is also possible that, in a particular objective-related module, a test may, in fact, consist of only one item. Clearly, when not all of the above data are available, not all of the decisions indicated in Table 1-1 can be made.

Observations from Table 1-1. Viewing our data in the manner indicated in Table 1-1 illustrates and reinforces the following observations.

(a) Decisions regarding instructional effectiveness necessitate a consideration of both pre- and posttest data. Decisions regarding mastery and/or grading involve a consideration of either pretest or posttest performance, but not both -- at least not in typical circumstances.

(b) Decisions regarding the efficiency of instruction necessitate a consideration of instructional effectiveness and the instructional time intervening between pre- and posttest. The importance of instructional time in learning has been treated by Carroll (1963, 1973); in fact, this issue is one of the primary motivating factors in Bloom's mastery learning model.

(c) Norm-referenced tests can serve a useful function in grading students. This author suggests that a student's grade be based on both norm-referenced and criterion-referenced information, since grades seem to be used as both a measure of what a student knows and as a measure of how much a student knows compared to what other students know.

(d) Decisions about validity and reliability are relevant for all kinds of tests. Furthermore, decisions about validity and reliability should be made for the "change" scores indicated by instructional effectiveness and retention.

These and other points concerning the issues raised by Table 1-1 are treated in much greater depth by Brennan (1973a).

CHAPTER II

Classical Test Theory in Criterion-Referenced Testing

Background

There seems to be some question in the minds of some researchers concerning the applicability of the classical test theory model to criterion-referenced testing. This is a potentially serious concern in that, if the assumptions of classical test theory are not met be criterion-referenced tests, then psychometricians evaluating such tests have virtually lost the benefit of over fifty years of test theory development. Even if criterion-referenced tests meet the assumptions of classical test theory, this is not a guarantee that the classical results and theorems form a sufficient theoretical basis for criterion-referenced tests; however, this problem is not nearly as serious as the problem of assumptions. Some aspects of the applicability of classical test theory have been discussed by Popham and Husek (1969), but they have not discussed the validity of the assumptions of classical test theory in relation to criterion-referenced tests. Therefore, it seems appropriate to analyze the assumptions of classical test theory in order to determine if any of these assumptions are not met by criterion-referenced tests.

In the author's opinion, the assumptions that we will discuss are often misunderstood or not fully appreciated. For example, many educators seem to have a virtual psychological fixation on the normal curve: Such educators tacitly assume that the normal curve is a sine qua non for classical test theory. As will be shownm however, this is not the case -- the assumptions of classical test theory are distribution free.

Notation

Unfortunately, there is no standard notation used by all writers who work in the field of classical test theory. The notation used by Gulliksen (1950) is simple, but not always sufficient; the notation used by Lord and Novick (1968) is very precise but perhaps more complicated than necessary for most researchers and practitioners. Therefore, an effort will be made to combine the most favorable aspects of both schemes of notation in the hope that the reader will be able to apply the adapted notation scheme to both of the above basis references.

Let X_{gi} be the observed score for the i-th person on test g, where K_g is the total number of items on test g and N is the total number of persons; i.e.,

$$X_{gi} = \sum_{j=1}^{K_g} u_{gij} ,$$

where u_{gij} is the score on item j of test g for person i.

Let T_{gi} be the true score for the i-th person on test g. The assumptions that will be stated later serve to define what we mean by "true score." One of the theorems that can be proved is that, in the classical test theory sense, the true score is the expected value of the observed score.

Let E_{gi} be the error score for the i-th person on test g. E_{gi}, called the "error of measurement," is the result of various chance or random factors that cause a person to answer correctly items he does not know or to answer incorrectly items he does know. Note that the errors accounted for by E_{gi} are chance errors, not systematic errors.

It is worth noting that T_{gi} is a fixed quantity for person i, but X_{gi} and E_{gi} are random variables. If the same person were given the same test a number of times, and if after each testing the person's "brains were washed" we would expect that the person's observed scores and error scores would show some variation; however, the person's true score is constant by definition.

Assumptions

The following assumptions express the posited relations between X_{gi}, T_{gi}, and E_{gi}.

A1: Definition of Random Error

$E_{gi} = X_{gi} - T_{gi}$ or $X_{gi} = T_{gi} + E_{gi}$;

A2: Zero Average Error

$\xi(E_{gi}) = 0$ in every non-null subpopulation of persons;

A3: Zero Correlation between True Score and Error Score

$\rho(T_{gi}, E_{gi}) = 0$;

A4: Definition of Parallel Tests -- parallel tests f,g, and h are defined as tests for which

(i) $T_{fi} = T_{gi} = T_{hi}$,

(ii) $\sigma^2(E_f) = \sigma^2(E_g) = \sigma^2(E_h)$, and

(iii) $\rho(T_{fi}, T_{gi}) = \rho(T_{fi}, T_{hi}) = \rho(T_{gi}, T_{hi})$;

A5: Zero Correlation between Errors on Parallel Tests

$\rho(E_{gi}, E_{hi}) = 0$ for parallel tests g and h;

A6: Zero Correlation between Errors on One Test and True Scores on a Parallel Test

$\rho(E_{gi}, T_{hi}) = 0$ for parallel tests g and h.

In the above assumptions A1 - A6, ξ indicates the expected value over persons in the subpopulation of persons under consideration, ρ indicates the correlation in the population of persons, and σ indicates the standard deviation in the population. The subscripts f, g, and h are reserved for parallel tests.

The set of assumptions A1 - A6 is actually more than sufficient. For example, Gulliksen (1950, pp. 6-13) does not list A4 (iii) and A6 as assumptions, since it is possible to prove both of these relations from the other assumptions.

Assumptions A1 - A6 are primarily based upon a consideration of the errors of measurement E_{gi}. It is also possible to state the above assumptions in terms of the true scores, but the assumptions then become somewhat more difficult to understand, and the resulting theorems necessitate more complicated derivations. Furthermore, we wish to concentrate upon errors of measurement since they form the crux of several arguments presented later in this chapter.

Let us now analyze the meaning of these assumptions.

Assumption A1 -- definition of random error. Assumption A1 postulates a linear relationship between the observed, true and error scores for a (randomly chose) person i. We are, in effect, saying that error score is the simple difference between true and observed score. Since, however, only X_{gi} is directly observable, the linear relationship contains two unknown quantities and is,therefore, undefined without additional information.

Assumption A2 -- zero average error. Assumption A2 states that given any non-null subpopulation of persons (where the population is countably infinite) the expected value of the error scores over persons is zero. In practice, the larger the number of cases in the distribution, the closer this assumption will be approximated. This assumption implies two important results:

(a) In the entire population of persons, the expected value of error scores is zero, and

(b) In every subpopulation consisting of persons with the same true score, the expected value of the error scores is zero.

The latter result may be written mathematically as:

$$\xi(E_{gi} | T_g) = 0 \;;$$

i.e., the expected value of E_{gi} for given true score T_g is zero, or the regression of error scores on true score is a horizontal straight line passing through the origin.

Assumptions A1 and A2 serve to define what is meant by true score. Also not that these assumptions imply that

$$\xi(X_{gi}) = T_{gi} \;.$$

Assumption A3 -- zero correlation between true scores and error scores. Assumption A3 states that the correlation between true and error scores in the population is zero. This means that we assume that there is no reason to expect positive (negative) errors to occur more frequently with high (low) true scores than with low (high) true scores. Note that Assumption A3 does not mean that error scores are distributed independently of true score. If errors are uncorrelated, "this merely means that the product-moment correlation is zero; if they are independent, this means that the frequency distribution of errors of measurement is the same regardless of the examinee's true score (Lord, 1959a, pp. 331)."

Assumption A4 -- definition of parallel tests. Assumption A4 serves to define what we mean by parallel tests. Parallel tests are tests that have (i) the same true scores, (ii) the same population variances, and (iii) identical intercorrelations. Of course, if there are only two parallel tests, then (iii) becomes meaningless. The concept of parallel tests may initially appear to be of secondary importance; however, parallel tests play an important role in classical test theory.

Assumption A5 -- zero correlation between errors on parallel tests. Assumption A5 states that the population correlation between random errors of measurement on parallel tests is zero.

Assumption A6 -- zero correlation between errors on one test and true scores on a parallel test. Assumption A6 states that population correlation between random error scores on one test and true scores on a second parallel test is zero.

Classical Assumptions and Criterion-Referenced Testing

We stated in passing that the assumptions of classical test theory are distribution-free, then we went on to discuss each of these assumptions. Note that none of the above assumptions necessitate any knowledge about the distribution of observed, true, or error scores. This fact, in addition to the very general nature of the assumptions themselves, seems to argue quite strongly that the classical test theory model is appropriate for criterion-referenced tests. At least, this author knows of no definition of criterion-referenced testing that, de facto, involves a violation of the classical test theory assumptions.

There is, however, at least one potential problem with the classical test theory assumptions in certain criterion-referenced and mastery testing situations. Consic r the subset of persons whose true score equals the highest possible true score for the test under consideration. From Assumption A2 we know that the expected value of the error scores for persons with the highest possible true score must be zero. In order for this to be true, positive and negative errors must be offsetting, but the highest possible true score equals the highest possible observed score. Therefore, all errors about the highest possible true score must always be zero. This conclusion seems difficult to support. It is somewhat analogous to saying that brilliant people are never subject to chance or random errors in their field of expertise.

Incidentally, this reservation about the classical test theory model is theoretically valid in norm-referenced testing situations as well as in criterion-referenced testing situations. However, in most norm-referenced testing situations the probability that a person will have the highest possible true score is rare; whereas this is not always true in criterion-referenced and mastery testing situations. Thus, the above reservation is potentially more serious for criterion-referenced tests than for norm-referenced tests. In either case, however, this author is not convinced that the reservation noted above is a devastatubg criticism of the classical model. Perhaps models can be posited that obviate this problem, but, in the meantime, the classical test theory model seems to provide a reasonable initial model for considering criterion-referenced tests.

It is useful to keep in mind three facts about the classical model: (a) $X = T + E$, (b) errors are random errors, and (c) a person's true score is the expected value of a large number of observed scores for that person. We wish to note one other aspect of the classical test theory assumptions. None of the assumptions necessitate that items be scored in the usual correct/incorrect manner. We will refer to this scoring procedure as the "classical scoring procedure"; however, the classical scoring procedure is not a necessary condition for classical test tneory.

Weak and Strong True Score Models

No matter how one applies the classical test theory model, it is clear that its assumptions do not constitute a very strong statistical framework for evaluating any test. In fact, the classical assumptions together with theorems that can be proven from these assumptions constitute a "weak" true score model, weak in the sense that all results are distribution-free (lord, 1965). However, it is a truism in mathematics that the weaker the assumptions, the weaker the results (Novick, 1966). In the last ten years, therefore, psychometricians such as Lord, Keats, and Novick have attempted to develop some strong true score models for tests. (Lord and Novick, 1968, is perhaps the best reference for the currently available strong true score models.) These models make stronger assumptions than A1 - A6 in the previous section, and the results that can be derived are likewise stronger.

Actually, one true score model (although it is seldom called a "true score model") has been in vogue for a considerable length of time. Many researchers (e.g., Gulliksen, 1950) have noted that in order to make use of the errors of measurement, it is necessary to make certain assumptions about the distribution of these errors (Gulliksen, 1950, p. 17). In the typical test theory situation it is usually assumed that these errors of measurement are normally distributed, independently of the true score, with mean zero in the population and constant population variance. It is primarily these assumptions that have led some naive users of classical test theory to mistakenly assume that classical test theory relates only to normally distributed test scores.

Characteristics of Errors of Measurement

It is instructive to consider the implications of only the classical assumptions upon errors of measurement. Assumptions A2 and A3 imply that

$$\xi(E_{gi} | T_g) = 0,$$

i.e., the regression of errors on true score is linear. More specifically, the best fitting line (in a least squares sense) is a horizontal straight line passing through the origin. Note that this does not mean that error scores are distributed independently of true scores; i.e., the distribution of error scores around any given true score is not necessarily the same. This means that the errors of measurement are unbiased; it does not mean that the variances of the errors of measurement around the true scores are equal.

Table 2-1 represents the observed scores and error scores for three true scores, where we assume that there are only three true scores and there are only as many people in the population as there are observed (or error) scores. Figure 2-1 represents the regression of these error scores on the true scores. The regression line is the line identical with the T-axis. Note that it is clearly true that $\xi(E_{gi}) = 0$ since $\xi(E_{gi} \mid T_g) = 0$ for every true score T_g. Since the regression line is horizontal, its slope is zero and consequently $\rho(T_{gi}, E_{gi})$ is also zero. Finally, note that the variances of the errors of measurement around the different true scores are not equal.

Now, it can be shown that

$$\xi(X_{gi} \mid T_g) = T_g$$

implying that the regression of observed scores on true scores is also linear (Lord and Novick, 1968, p. 65). Moreover, this regression line passes through the origin and its slope is equal to unity. Figure 2-2 shows a graph of this regression line for the data given in Table 2-1.

Neither one of the above regressions (represented by Figures 2-1 and 2-2, respectively) is however, the primary regression of interest. The test evaluator is usually primarily concerned about $\xi(T_{gi} \mid X_g)$, the regression of true scores on observed scores, in order to estimate a student's true score from his observed score. However, this regression can be non-linear, and consequently neither the true score nor the observed score distribution is necessarily normal (Lord and Novick, 1968, pp. 500-505).

Figure 2-3 shows a plot of the distribution of true scores (ordinate) versus the distribution of observed scores (abscissa) for the data in Table 2-1. The six circled points in Figure 2-3 represent three sets of (two) observed scores that map into different true scores. For example, according to Figure 2-3 an observed score of 12 can indicate a true score of either 10 or 20.

Table 2-1

The Relation Between True, Observed, and

Error Scores (Synthetic Data)

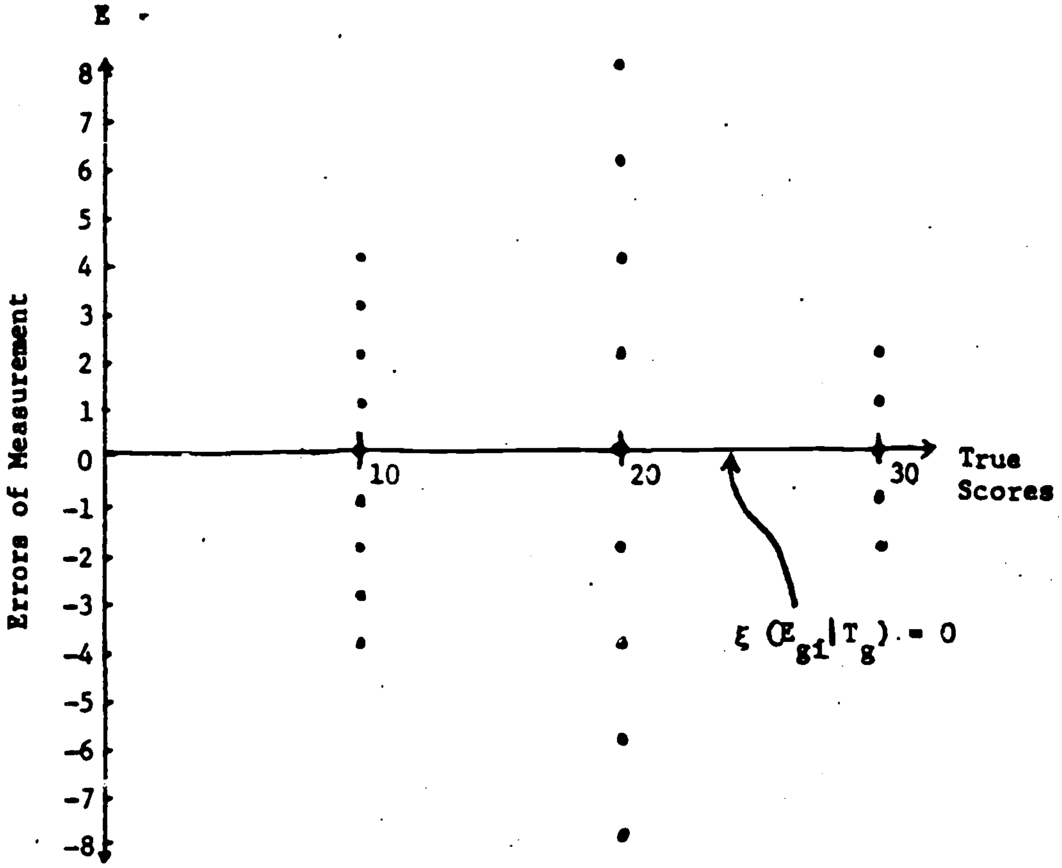
True score T_g	= Observed score X_{gi}	- Error score E_{gi}	$\xi \left(E_{gi} \vert T_g\right)$	$\xi \left(X_{gi} \vert T_g\right)$
10	6	-4	0	10
10	7	-3	0	10
10	8	-2	0	10
10	9	-1	0	10
10	10	0	0	10
10	11	1	0	10
10	12	2	0	10
10	13	3	0	10
10	14	4	0	10
20	12	-8	0	20
20	14	-6	0	20
20	16	-4	0	20
20	18	-2	0	20
20	20	0	0	20
20	22	2	0	20
20	24	4	0	20
20	26	6	0	20
20	28	8	0	20
30	28	-2	0	30
30	29	-1	0	30
30	30	0	0	30
30	31	1	0	30
30	32	2	0	30

Figure 2-1

The Regression of Error

Scores on True Scores

$\xi(E_{gi}|T_g) = 0$

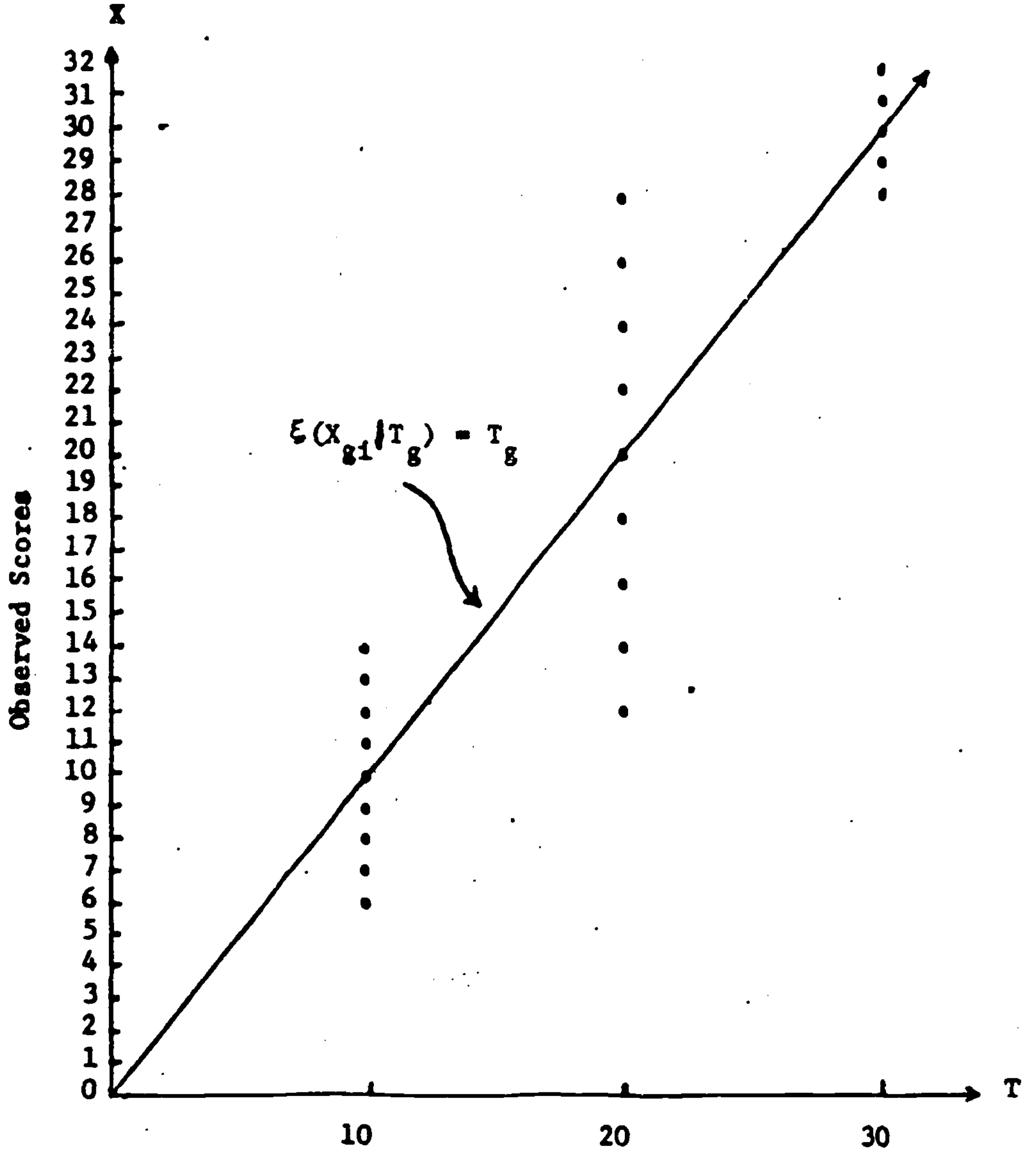


Note.--The data for the above figure are given in Table 2-1. The abscissa represents true score T and the ordinate represents error score E. Note that the variances of the errors of measurement about each of the true scores are not equal.

Figure 2-2

The Regression of Observed Scores on True Scores

$\xi(X_{gi}|T_g) = T_g$

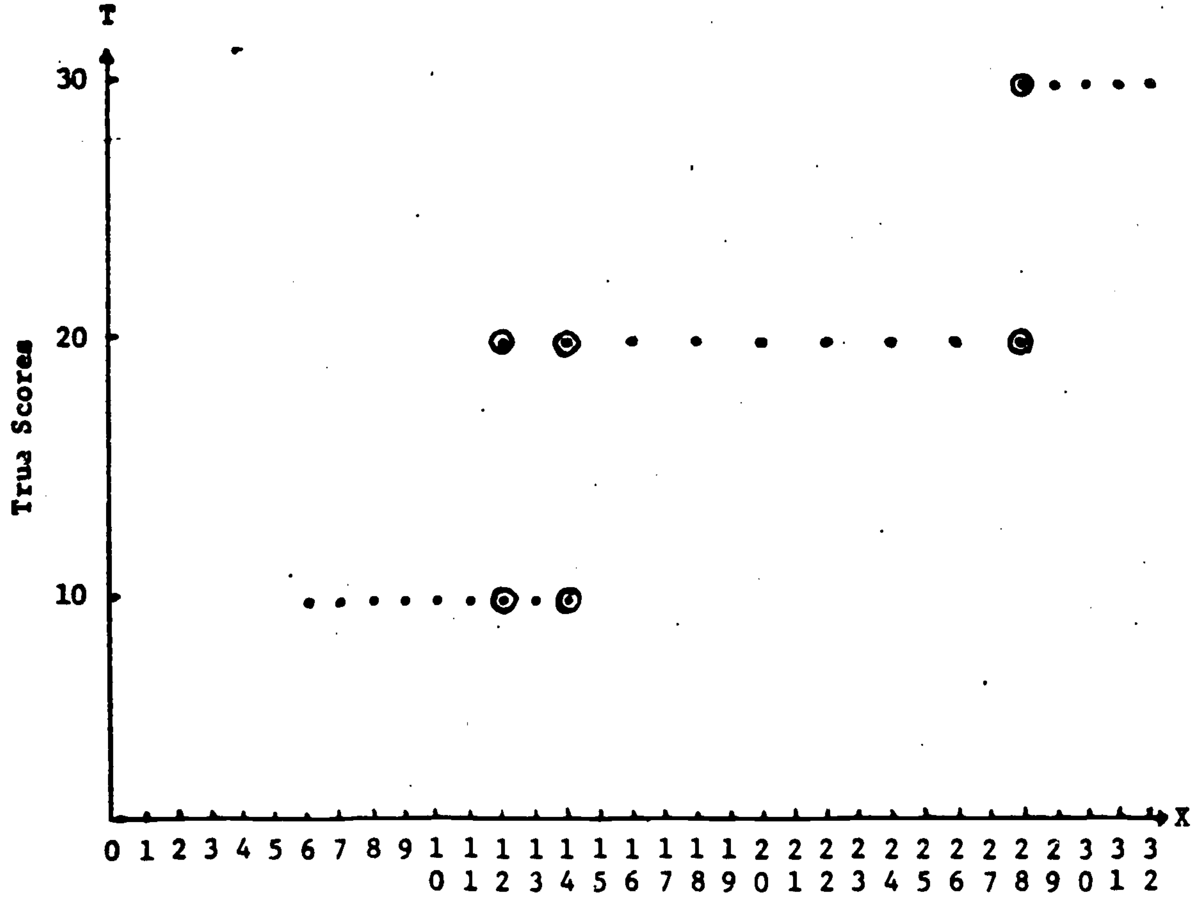


Note.--The data for the above figure are given in Table 2-1. The abscissa represents true score T and the ordinate represents observed score X. Note that the variances of the errors of measurement about each true score are not equal.

Figure 2-3

The Regression of True Scores on Observed Scores

$\xi(T_{gi} | X_g) = ?$



Note.--The data for the above figure are given in Table 2-1. The ordinate represents true score T and the abscissa represents observed score X. The circled points represent three sets (pairs) of observed scores that map into different true scores. For example, an observed score of 12 can indicate a true score of 10 or 20.

A similar situation occurs for observed scores of 14 and 28. The fact that certain observed scores do not map into unique true scores indicates that $\xi(T_{gi}|\ X_g)$ is not linear. If this is not clear, consider the following:

$$\xi(T_{gi}|\ 6 \leq X_g \leq 11) = 10\ ,$$

$$\xi(T_{gi}|\ X_g = 12,\ 14) = 15\ ,$$

$$\xi(T_{gi}|\ X_g = 13) = 10\ ,$$

$$\xi(T_{gi}|\ 16 \leq X_g \leq 26) = 20\ ,$$

$$\xi(T_{gi}|\ X_g = 28) = 25, \text{ and}$$

$$\xi(T_{gi}|\ 29 \leq X_g \leq 32) = 30\ .$$

The above expectations certainly do not constitute a linear function. A linear best-fitting regression line could be forced to fit thedata, but this would not be "the" best-fitting curve for the data; "the" best-fitting curve would be curvilinear.

Normal Error Model

The results shown in Figures 2-1, 2-2, and 2-3 are based only upon the assumptions of classical test theory. Consequently, these results are distribution-free. Note expecially that these results do not make any assumptions about the distribution of the errors of measurement. Gulliksen (1950, p. 17) notes that in order to make use of the errors of measurement we must make some assumptions about the frequency distribution of these errors. The assumptions that we will now discuss form the rationale behind the theory for norm-referenced tests that rely upon normally distributed true and observed score distributions.

When, in addition to the assumptions of classical test theory, we assume that (a) the errors of measurement are distributed independently of true score, (b) the errors of measurement are distributed normally with mean zero and constant variance, and (c) the regression of true scores on observed scores is linear, then both the true and observed scores must be normally distributed (Lord and Novick, 1968, p. 503) with

$$\sigma_T^2 = \sigma_X^2 - \sigma_E^2$$

In practice, the lastassumption (the linearity of the regression of true on observed scores) is often neglected (Gulliksen, 1950, Section 2.11), without serious difficulty. It can be shown that the linear regression of true scores on observed scores is given by

$$\xi(T_{gi} \mid X_g) = (1 - \rho_{XX})\mu_X + \rho_{XX}X_g \text{ , where}$$

$$\rho_{XX} = \beta_{TX} = \sigma_T^2 / \sigma_X^2 \quad \text{or}$$

$$\rho_{XX} = \rho_{XT}^2$$

In both cases ρ_{XX} is called the reliability coefficient (which is also the correlation between parallel measurements).

Assumptions (a) and (b) are indicated in Figure 2-4. Note the difference between the distributions of the errors of measurement indicated in Figures 2-1 and 2-4. In both figures, $\xi(E_{gi} \mid T_g) = 0$, but in Figure 2-4 the errors of measurement have a specific distributional form (i.e., the same normal distribution) for each true score T_g.

Assumptions (a), (b), and (c) above thus provide the basis for a strong true score theory of test scores that results in normally distributed true and observed scores. In practive, these assumptions are the ones most frequently made (either consciously or unconsciously) about errors of measurement. These are the assumptions that make it possible to evaluate and interpret most of the currently available norm-referenced tests.

The Relevance of Normality Assumptions to Criterion-Referenced Tests

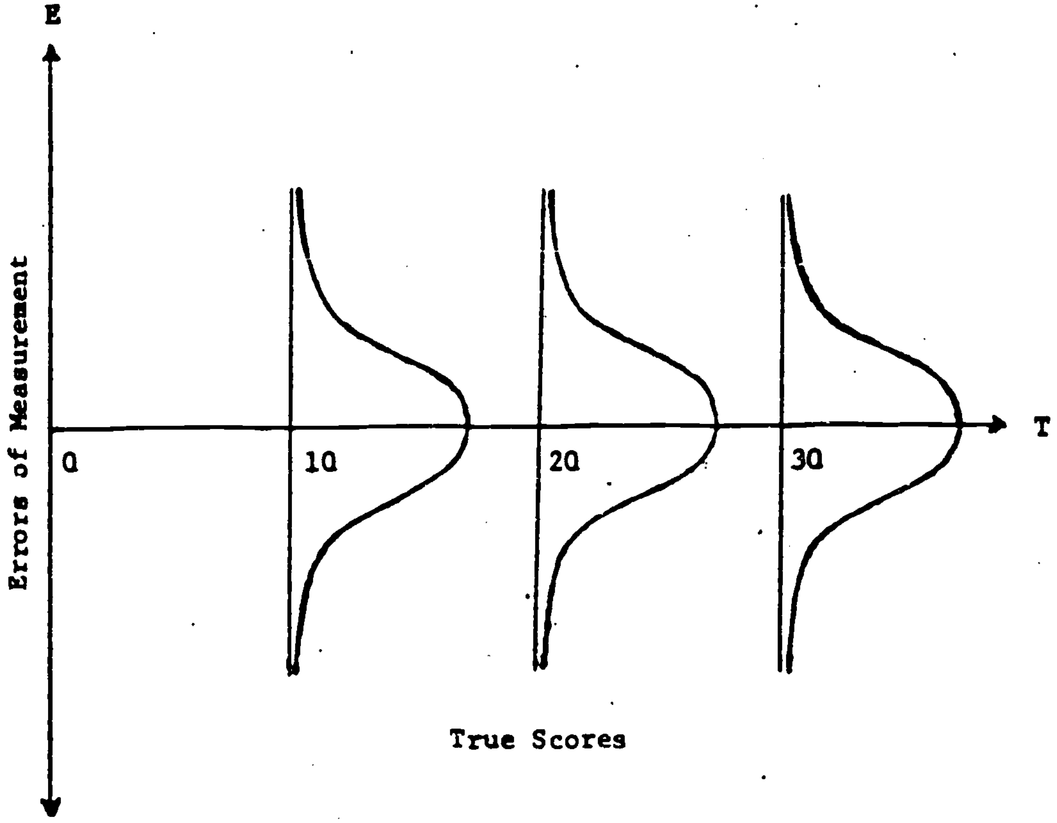
There is no doubt that the normality assumptions presented in the previous section are very useful for many testing purposes; however, these assumptions do not seem to be applicable for many criterion-referenced testing situations. Lord states:

> The assumption that each error is distributed $N(0,\sigma^2)$ independently of true score is probably quite adequate for many purposes. However, it is clear that these assumptions cannot be met when the

Figure 2-4

The Assumption of N(0, σ_E^2) Distributed

Errors of Measurement



Note.—In the above figure we assume that there are only three possible true scores on a (hypothetical) test. The errors of measurement about each true score are normally distributed with mean zero and constant variance σ_E^2.

> true score, expressed as a proportion of the number of items in a test, is near zero or near one If n is the number of test items, and τ/n is some small number like .01, it is intuitively obvious, in view of the fact that the observed test score can never be negative, that the distribution of the errors of measurement will in all probability be skew, and that the standard deviation of this distribution will surely be less than if the true score were not so near to zero (Lord, 1959b, p. 475).

Similarly, if the true score expressed as a proportion of the number of items correct is near unity, then the distribution of the errors of measurement will be less than if the true scores were not so close to unity. If, in either case, it were assumed that the errors of measurement were distributed independently of true score with constant variance about each true score, this would imply that certain (postulated) observed scores would, in fact, be unobtainable. (See Figure 2-5.) Lord (1960) discusses in some depth the consequences of assuming that errors of measurement are distributed

$$N(0,\sigma_E^2)$$

independently of true score.

Since, for many criterion-referenced tests, many of the students get most of the items correct, it is obvious that we often expect the true proportion of items correct for at least some student to be near unity. Thus, on the basis of the arguments presented above, it should be clear that the normality, constant variance, and independence assumptions presented in the previous section are not always applicable for criterion-referenced tests. The next section describes assumptions that are, however, quite appropriate for such tests.

The Binomial Error Model

Recall that the normal error model assumes that (a) the errors of measurement are distributed independently of true score and (b) the errors of measurement are distributed normally with mean zero and constant variance for each true score, i.e.,

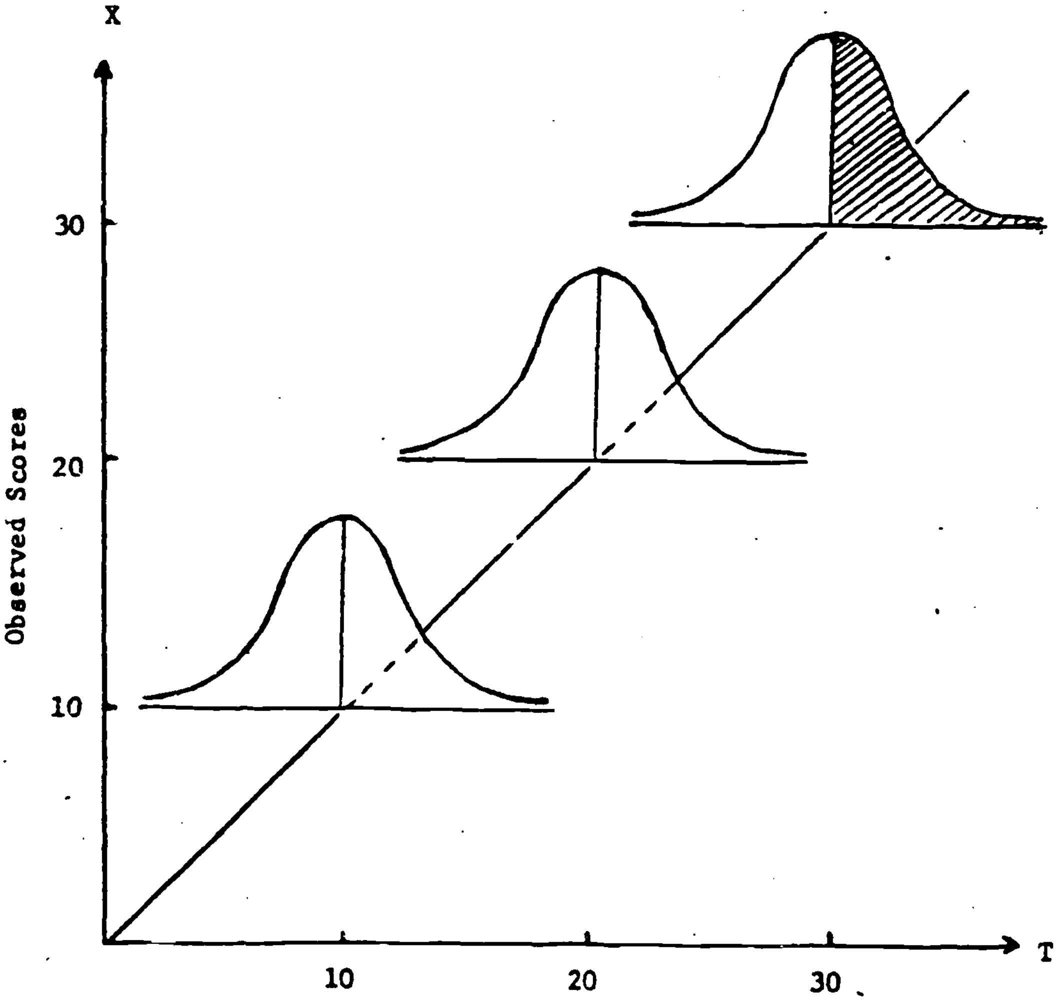
$$N(0, \sigma_E^2) .$$

In the previous section we demonstrated that neither (a) nor (b) is reasonable for at least some criterion-referenced tests. Thus, for such tests we are forced to

Figure 2-5

A Consequence of Assuming N(0, σ_E^2) Distributed

Errors of Measurement



Note.--The shaded area indicates scores that are not obtainable assuming that the maximum score on the test is 30.

make assumptions about the e᷊rors of measurement. A little over a decade ago Keats and Lord (1962) postulated that a reasonable distributional form for the errors of measurement (assuming that observed scores are bounded) is the binomial distribution with its parameter equal to a specified true score. Keats and Lord (1962) give no indication that they were, at that time, even considering what we now call criterion-referenced tests; however, as will be demonstrated, the implications of this assumption correspond quite well to a working definition of the distributional form of many criterion-referenced test scores.

More extensive discussions of the binomial error model can be found in Keats and Lord (1962), Keats (1964), Lord (1965), Lord and Novick (1968) and Brennan (1970). The last reference is intended to provide a simplified and concise description of the binomial error model, especially for those interested in its possible application in criterion-referenced testing. In this report we will merely provide a brief outline of the binomial error model.

As far as notation is concerned, subscripts for variables will be dropped unless they are required to avoid ambiguity. As before, T represents true score, X represents observed score, and E represents error of measurement. However, rather than T, the true score number of items correct, we will be concerned primarily with ζ, the true score proportion correct; i.e.,

$$\zeta = T/N$$

where N is the number of items on the test. Note that the observed score, X, is a discrete variable, while the true score, ζ(as well as T), is assumed to be continuous. Distributions will be identified as follows:

$\phi(X)$ = the distribution of observed scores,

$g(\zeta)$ = the distribution of true scores,

$f(E|\zeta)$ = the distribution of the errors of measurement for given true score , and

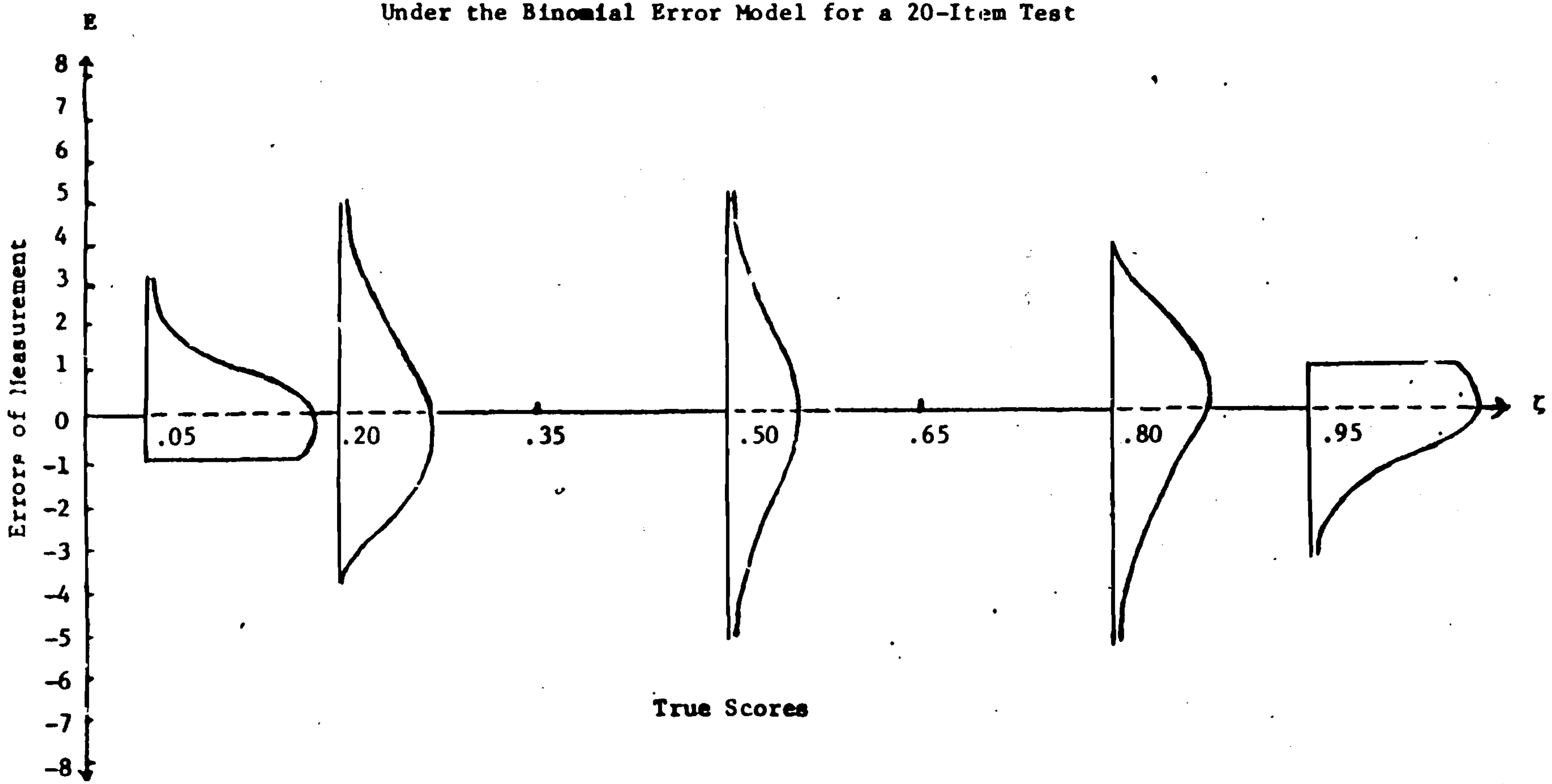
$h(X|\zeta)$ = the conditional distribution of observed scores for given true score.

Recall that the binomial error model is a strong true score model; i.e., it incorporates an assumption(s) over and above the assumptions of the classical weak

Figure 2-6

The Conditional Distribution of Errors of Measurement for Several True Scores

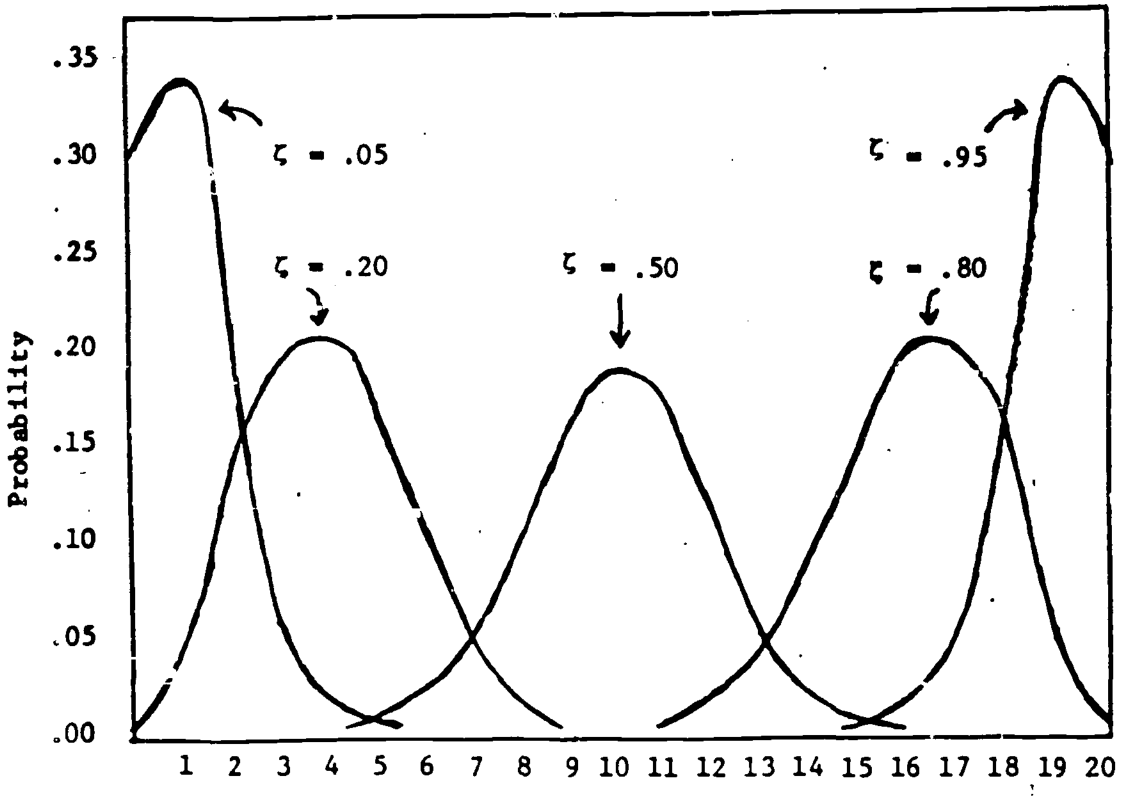
Under the Binomial Error Model for a 20-Item Test



Note.—The numbers on the ζ (true score) axis represent the true score proportion of items correct; i.e., T = 20ζ , where T is the true score in terms of number of items correct.

Figure 2-7

The Conditional Distribution of Observed Scores for Several Given True Scores under the Binomial Error Model for a 20-Item Test



Number of Items Correct
(Observed Scores)

Note.--The above figure represents five different observed score distributions determined by the parameter ζ . This figure represents the same information as that contained in Figure 2-6.

true score model. The binomial error model does not, therefore, violate any of these classical assumptions A1 -A6; all these assumptions still hold.

In addition, however, for the binomial error model, it is assumed that for a given true score, ζ, the errors of measurement, E are independent and have a binomial distribution with parameter ; i.e.,

$$f(E|\zeta) = \binom{N}{X} \zeta^{X}(1 - \zeta)^{N-X} \quad , \quad X = 0,1, \ldots, N$$

for given true score ζ, where N is the number of test items. This assumption can be stated as follows: the conditional distribution of observed score X for given true score ζ is the binomial distribution with parameter ζ: i.e.,

$$h(X|\zeta) = \binom{N}{X} \zeta^{X}(1 - \zeta)^{N-X} \quad , \quad X = 0,1, \ldots, N.$$

The first of these two formulas is illustrated for a 20-item test in Figure 2-6. It is instructive to compare Figure 2-6 with Figures 2-1 and 2-4, which illustrate the error assumptions for the classical model and the "normal" model, respectively. The second of these two formulas is illustrated in Figure 2-7.

Mathematically, assuming a linear regression of true scores on observed scores, the above assumptions imply that observed scores have a hypergeometric distribution and true scores have a beta distribution. Both of these distributions can take on the negatively skewed characteristic of many criterion-referenced and mastery test score distributions.

Several times in the above discussions we have assumed that the regression of true scores on observed scores is linear. We also mentionned that this assumption is not always true; however, Lord and Novick (1968) claim that departures from linearily are probably not too great, in most cases. This is one reason we have stuck with the linearity assumption. Another reason is that any non-linear assumption about the regression of true on observed scores would necessitate relatively complicated calculations in order to determine the regression equation, the distribution of observed scores, and the distribution of true scores. A third reason is that it seems wise to consistently assume the linearity of the regression of true on observed scores so that the reader can more effectively compare the test theory models discussed.

A fairly up-to-date and extensive discussion of applications of the binomial error model (for settings not necessarily related to criterion-referenced tests) is given in Lord (1965). A further extension of the binomial error model (called the "compoind binomial error model") is given in Lord and Novick (1968).

Summary and Discussion

We have reviewed the classical test theory model and found it to be generally applicable to criterion-referenced testing with two reservations: (a) there is some doubt about the applicability of the model for the subset of persons who have the highest possible true score and (b) the model may be appropriate, but not sufficient, for criterion-referenced testing.

Also, we have reviewed the implication of the classical test theory assumptions upon errors of measurement, and we have reviewed the normal and binomial error models. We find that, for criterion-referenced testing, if the classical model is to be used, then the binomial error model assumptions are more appropriate than the normal error model assumptions, in most cases. However, we should note in passing that it is considerably more difficult and time-consuming to use the binomial error model than to use the normal error model.

In the context of this chapter, the normal and binomial error models provide us with alternative ways to estimate a person's true score given that person's observed score. Our discussion of error scores and their distribution is not necessarily appropriate for considering whether a student is above or below a mastery cutting score. Hambleton and Novick (1973) seem to consider this issue to be the crucial issue in criterion-referenced testing. It seems to me that whether or not a person is above or below a mastery cutting score is critical in mastery testing, but may not be critical for criterion-referenced testing, in general. Recall that a criterion-referenced test must have a "specified performance standard," but this "standard" does not necessarily require a mastery cutting score.

In any case, whether one is dealing with mastery testing or its progenitor, criterion-referenced testing, it seems to this author that the estimation of a student's true score is a critical consideration. If one assumes the classical definition of true score, then the binomial error model seems appropriate; if one assumes a definition of true score that depends upon a mastery cutting score, then Hambleton and Novick's (1973) suggestions seem reasonable, but even their suggestions depend indirectly upon the classical definition of true score; and, finally if one assumes a different definition of true score, then different assumptions about errors may be necessitated.

In conclusion, our discussion of the distributions of different kinds of scores may seem inconsistent with previous statements about the irrelevance of score distributions in criterion-referenced testing. To be more specific, as far as the interpretstion of a set of criterion-referenced test scores is concerned, the distributions of observed and true scores over persons are irrelevant; however, assuming that one wants to estimate a given person's true score, one must make assumptions about the distribution of errors of measurement for that person, or, more specifically, one must make assumptions about the distribution of errors of measurement for all persons who have an observed score equal to the given person's observed score. Therefore, the distribution of errors of measurement is critical in criterion-referenced testing, eventhough the distributions of observed and true scores over persons are not relevant. It just so happens that a unique description of the distributions of true and observed scores is a by-product of assuming (a) the classical test theory model, (b) the linearity of the regression of true scores on observed scores, and (c) either the normal or the binomial error model.

CHAPTER III

Validity and Procedures for Constructing Criterion-Referenced Tests

The most important aspect of any test is its validity; i.e., the extent towhich the test measures what it is intended to measure. For criterion-referenced tests, most researchers view the question of validity primarily as a question of content validity (see Popham and Husek, 1969). For the most part, this author agrees with this view. However, our concern for content validity argues that we also consider the most important procedures involved in constructing criterion-referenced tests. It is these procedures that provide a basis for inferring the extent to which a test has content validity.

For our purposes, let us consider five steps in the developemnt of criterion-referenced tests: (a) the establishment of a domain ofrelevant behaviors, (b) the development of a procedure to generate items, (c) the development of an item sampling plan, (d) the development of a procedure to administer items, and (e) the collection of data and the revision of the test. In the following sections, we will treat important aspects of each of these issues and provide major references for the reader interested in more detail. Many of the issues discussed in this chapter and the next two chapters are also treated from a somewhat different point of view by Rovinelli and Hambleton, 1973.

The Development of Criterion-Referenced Tests

Domain of relevant behaviors. The first step in the development of a criterion-referenced test entails specifying and categorizing all of the behaviors which are to be tested. Operationally, this frequently means specifying and categorizing a set of objectives; thus, the task is analogous to constructing a blueprint for a norm-referenced test. A more specific approach to the task of establishing (and using) a domain of relevant of behaviors is called "domain-referenced achievement testing"; Hively et al., 1973, provide an excellent statement of this model.

Two questions usually arise when one attempts to specify the domain of relevant behaviors: (a) how extensive should the domain be? and (b) what is the nature of the domain? This author knows of no generally accepted procedure for defining the extent of the domain. Frequently, the extent of the domain corresponds with the extent of the subject matter to be covered in a certain course, in a particular segment of instruction, or in a particular time period. From a measurement point of view, it is probably advisable that a domain, or each unambiguously defined subset of the domain, contain or reference a set of objectives which is tested by a single criterion-referenced test. If this advice is followed, then, of course, the objectives in a domain, or each subset of the domain, should be closely related. When these conditions prevail, the items in a given criterion-referenced test will be testing similar objectives, and, therefore, the interpretability of a criterion-referenced test score will, in general, be enhanced.

The nature of the domain may be considered as the way in which the objectives or elements of the domain are inter-related. When viewed in this manner, we can say that a domain can be characterized by: (a) no hierarchy, (b) a linear hierarchy, or (c) a complex hierarchy (i.e., a hierarchy having different branches). Now, one can postulate learning hierarchies (i.e., hierarchies indicating an optimum or desired order in which objectives should be taught to students in order to maximize learning) or knowledge hierarchies (i.e., hierarchies indicating which objectives are logical pre-requisites to attaining other objectives). In these terms, the hierarchy in our "domain of relevant behaviors" is typically a knowledge hierarchy, which need not necessarily correspond to a learning hierarchy for the subject matter under consideration. It should be noted that it is not always necessary to specify a knowledge hierarchy even if one exists or can be postulated. However, a knowledge hierarchy is at least useful and often essential when one undertakes sophistocated procedures for item sampling and/or item administration (see discussion below).

Procedures to generate items. At the present time, there are fundamentally two procedures for the generation of criterion-referenced test items: (a) have content specialists write items and (b) use item forms.

In many areas of education, the item forms approach is not feasible, from a practical point of view, at this time. Consequently, one must have content specialists write test items in these areas. There are, however, problems in having content specialists write test items. In fact, most of the typical issues that surround content validity emanate from a consideration of whether or not the items written by content specialists are unambiguous measures of intended objectives at the intended level of difficulty. It is especially difficult for content specialists to write "equivalent" test items for a given objective, and this is frequently a large part of the item writing task for criterion-referenced testing.

During the last few years, an excellent theoretical foundation for item generation has been provided by literature on "item forms," a term originally introduced by Hively (1962). Item forms make it possible to define an entire class of items merely by substituting elements of replacement sets for variable elements in the item forms. There are at least two important advantages of this item generation technique: (a) item forms provide a concrete basis for generalization to a domain of content, thus providing a sound basis for examining content validity, and (b) item forms allow for the possibility of generating a large number of equivalent items. In addition, Nitko (1970) argues that the analysis of a content area through the use of item forms provides a sound basis for the "systematic study of the domain of instructionally relevant tasks in terms of its structural and behavioral parameters (p. 10)."

The literature indicates basically three approaches to the construction of item forms. Hively et al (1968) and Ferguson (1969, 1971) use item forms primarily characterized by numerical replacement sets; Osburn (1968) uses item forms that employ both numerical and non-numerical replacement sets; and Bormuth (1970) argues for the use of item forms that incorporate linguistic transformational rules. All of the above researchers have, to varying degrees, treated the computer generation of test items through the application of item forms. Perhaps one of the best examples is provided by Ferguson (1969).

Item sampling. In considering item sampling in the context of criterion-referenced tests, it is useful to recall that: (a) each objective has at least one and usually many "equivalent" test items associated with it and (b) the domain consists of a set of possibly interrelated objectives. Now in choosing items for a criterion-referenced test there are a number of possible sampling schemes that might be employed. For example, the test might consist of : (a) all items, (b) a random sample of items, (c) a stratified random sample of items where stratification occurs with regard to objectives, or (d) a representative sample of items. Kriewall (1969) and Lord and Novick (1968) provide a partial consideration of these sampling plans.

A stratified random sample is perhaps the most common sampling method in criterion-referenced testing; however, the exact nature of the sampling plan and the sampling fractions are, unfortunately, seldom specified in detail. One reason that stratified random sampling is so popular is that it is practically ideally suited to the item forms approach to the generation of test items. Often, the item forms are the strata, and each item form provides a method of generating a set of items from which a random sample is drawn.

Item administration. Often, a criterion-referenced test is, as are most norm-referenced tests, a fixed entity; i.e., for each person taking the test, the items are the same, and the order in which the items appear on the test is the same for all persons. Sometimes the order of administering items varies for each student, or for several sets of students. Less frequently, different students are given different items; in such cases, it also frequently occurs that different students receive different numbers of items which may even come from different strata. This last kind of item administration technique can be referred to generically as "adaptive testing." A specific kind of adaptive testing is called sequential testing, the statistical aspects of which are treated by Wald (1947). More recently Kriewall and Hirsch (1969) consider sequential testing in the context of criterion-referenced testing.

Adaptive testing (see Brennan, 1973) has been employed in both criterion-referenced and norm-referenced testing situations. This testing technique has been called "tailored testing" (Lord, 1971), "branched testing" (Ferguson, 1969, 1971), "programmed testing" (Linn et al, 1969) and "sequential testing" (Linn et al, 1970). These types of tests have some elements in common; however, there are often differences among the ways these terms are used by particular researchers. Therefore, we will group all of the above testing techniques under the general headimg of "adaptive testing," as distinct from "conventional testing" in which all items are administered to all examinees usually during a fixed-length time period.

In general, current research on adaptive testing can be divided into two types: (a) adaptive testing for norm-referenced testing and (b) adaptive testing for criterion-referenced testing. Lord (1970, 1971) examines relevant issues in norm-referenced adaptive testing, while Ferguson's (1969, 1970) research treats criterion-referenced adaptive testing. Linn et al (1969, 1970) incorporate aspects of both types of adaptive testing, although they seem to view the achievement testing process primarily from a norm-referenced viewpoint.

Norm-referenced adaptive testing involves tailoring item difficulties to examinees in such a way that examinees spend most of their time answering items at or near their ability level. The objective is to determine an individual's relative potition on some hypothetical continuum of underlying ability. Unfortunately, theoretical work by Lord (1970) indicates that the types of norm-referenced adaptive tests thus far examined have serious limitations -- they do not provide "greatly improved measurements for most examinees. The value of (these) tests is primarily for those examinees for whom the conventional test would be too easy or too difficult (Lord, 1970, p. 153)." Thus, at this time, it appears that adaptive testing offers no significant advantages for the conventional types of norm-referenced tests. It is interesting to note, however, that most of the above research is of a theoretical nature; in practice, the computerized administration of such tests might yield significant improvements in reliability and validity per unit of testing time.

In any case, testing within the context of instruction typically involves a different kind of measurement from that discussed by Lord (1970, 1971). In instruction, it seems more appropriate, in most cases, to employ criterion-referenced measurement instruments in such a way that decisions can be made concerning whether or not each student has achieved a desired level of proficiency.

In the opinion of this author, the best example of criterion-referenced adaptive testing for instructional decision-making is provided by Ferguson (1969). He postulated a knowledge hierarchy for elementary addition and subtraction problems and developed a computerized system that employs the theory of item forms to generate a set of criterion-referenced test items for each of the nodes of the hierarchy. Then, using the sequential probability ratio test (Wald, 1947) as a primary basis for decision-making, he created an adaptive test which, "when compared to ... conventional tests (for determining proficiency in elementary addition and subtraction) ... seems comparable or superior in all respects (Ferguson, 1969, p. 88)." Furthermore, his test was effective and efficient for determining the proficiency of all examinees, even those in the middle range of proficiency.

Thus, while the research findings for adaptive testing in the norm-referenced context are less than promising, the findings for criterion-referenced adaptive testing in the context of instructional decision-making are quite encouraging.

Data collection and test revision. In a subsequent chapter, we discuss the role of empirical data in the revision of test items. Here, we merely outline several issues of general importance.

Ideally one should collect and analyze data from all subjects who take the test in order to identify test items, and other aspects of the test, that require revision. If this is not feasible, one can analyze data from a random sample or representative sample of subjects; however, one must be careful to obtain a large enough sample so that item statistics are reasonably stable. The experience of this author indicates that the minimum sample size should be about 25-30 subjects, if at all possible.

A second consideration is that empirical data should not form the sole basis for the revision of a criterion-referenced test. Data may indicate the _potential_ need for revision; however, whether or not revision is actually undertaken should ultimately depend upon the judgment of subject matter specialists who have studied the data and weighed the trade-offs involved in revision.

A third consideration is that the process of creating a good criterion-referenced test (or any test, for that matter) is a cyclic process of replication and revision. If the revision process employed is adequate, then each revision of the test should be an improvement of the previous version. This last statement may appear trivial; however, it should be noted that not all revision procedures necessarily result in improvement.

Some Issues Concerning the Validity of Criterion-Referenced Tests

We mentioned at the beginning of this chapter that content validity is a principal concern for criterion-reference^d tests. If the procedures indicated above are followed carefully, then one has a reasonable expectation of obtaining a criterion-referenced test having content validity. Perhaps the most crucial issue is what Dahl (1971) calls "objective-item congruence", i.e., the extent to which the criterion-referenced items are appropriate measures of the objectives in the domain of intended behaviors.

The critical nature of objective-item congruence provides, I think, an important argument in favor of the item forms approach to the generation of test items. The item form is usually an operational definition of the objective, and the item form provides a basis for generating the test items for the objective; hence, one has a strong logical basis for arguing that the test has content validity (in the sense of objective-item congruence) when one uses the item forms approach to the generation of test items. There is, however, one caution that should be noted concerning the use of item forms -- the items resulting from a particular item form are not necessarily equivalent in a statistical sense. For example, it is not necessarily true that items generated from the same item form will all have the same (empirical) difficulty level.

When subject matter specialists generate items, it becomes necessary to employ some judgmental procedures in order to assess the content validity of the criterion-referenced test. Such judgmental procedures usually entail assessing the extent to which subject matter specialists, working independently, agree that the test has objective-item congruence. There are a number of procedures for assessing agreement between or among judges. The reader may be interested in referring to Light (1973) for an excellent review of the literature in this area. Three potentially useful techniques, in the opinion of this author, have been discussed by Lu (1971), Hemphill and Westie (1950), and Brennan and Light (1973).

Another issue relating to the question of validity involves the nature of the student scores on the test. A criterion-referenced test, by definition, necessitates "measurements that are directly interpretable in terms of specified performance standards (Glaser and Nitko, 1971)." Therefore, the extent to which a criterion-referenced test is valid depends upon the extent to which scores reported on such a test are interpretable in terms of specified performance standards. For example, one may be concerned about the proportion of items, from a given domain, to which a student knows the answer. In this case, a student's score is valid to the extent that the observed proportion is free of random and systematic errors of measurement. Or one may be concerned about whether or not the true (in the sense of "actual") proportion of items to which the student knows the answer is above or below some mastery cutting score. In this case, students' scores are valid to the extent that both random and systematic errors of classification (of students above and below the mastery cutting score) are eliminated.

The above observations point to a central relationship between reliability and validity for criterion-referenced tests (or any test, for that matter). This relationship may be stated as follows: a test is reliable to the extent that scores resulting from it are free of random errors of measurement; a test is valid to the extent that scores resulting from it are free of both random and systematic errors of measurement.

CHAPTER IV

Reliability of Criterion-Referenced Test Scores

The most frequently considered statistical issue surrounding criterion-referenced measurement involves the reliability of such measures. In this chapter, we review fundamental ideas about reliability, we consider several problems in employing norm-referenced reliability measures for criterion-referenced tests, and we provide a critical review of most of the reliability measures that have been suggested for criterion-referenced tests.

Classical Notions about Reliability

In the classical test theory model, reliability is defined as either (a) the squared correlation between true scores and observed scores or (b) the ratio of the variance of true scores to the variance of observed scores. (Gulliksen, 1950, and Lord and Novick, 1968, treat the theory of reliability in considerable detail.) Neither of these two theoretical definitions of reliability can be applied directly since they involve the unobservable true scores discussed in Chapter II.

However, under the classical test theory model it can be shown that reliability is also equal to the correlation between parallel tests, where parallel tests are defined statistically as tests that have equal means, equal variances, and equal intercorrelations (if there are more than two tests involved). Therefore, one method of determining reliability is to obtain the correlation over persons on parallel tests; this is called a measure of equivalence.

Another measure of reliability is called a measure of stability, which is the correlation over persons of two separate administrations of the same test, with the assumption that no learning occurs between the first and second administrations.

A third measure of reliability is called internal consistency. Typical measures of internal consistency are Kuder and Richardson's (1937) Formulas 20 and 21, Cronbach's Coefficient Alpha (1951), and Hoyt's (1941) Reliability Coefficient. For the classical correct/wrong scoring procedure these coefficients (with the exception of Formula 21) all provide identical results.

O'her kinds of internal consistency measures include measures of homogeneity and split-halves coefficients. Some authors consider measures of internal consistency as different from measures of reliability (e.g., Brown, 1970); most authors, however, treat measures of internal consistency as a special kind of measure of equivalence. A critical point to recognize is that measures of internal consistency employ only one administration of one test.

A measure of reliability in and of itself is essentially a statistic characterizing the extent to which a test is a dependable measurement instrument. However, indirectly a reliability coefficient provides a basis for making inferential statements about true scores and observed scores. (See discussion of errors of measurement in Chapter II.)

Also, although we usually consider reliability as a measure involving a test of fixed length, one can use the Spearman-Brown Prophecy Formula to estimate the reliability of a test of any length. An important special case is the reliability of a one-item test, which is mathematically equal to the intraclass correlation coefficient for the test of full length. The reader interested in new and important developments concerning these and other related issues should consult Cronbach et al (1972).

Another important issue in reliability theory involves the reliability of change scores. (See, for example, Harris, 1963, Tucker et al, 1966, and Cronbach et al, 1970). For example, one typically judges the effectiveness of an instructional system in terms of pretest-posttest changes in student performance. Therefore, in order to judge the effectiveness of an instructional system one needs to know the reliability of these change scores. This is a very complicated issue and one that has not received a great deal of treatment from a criterion-referenced testing viewpoint.

Problems in Using Norm-Referenced Reliability Indices for Criterion-Referenced Tests

A few years ago Popham and Husek (1969) stated:

> "... it is obvious that a criterion-referenced test should be internally consistent. If we argue that the items are tied to a criterion, then certainly the items should be quite similar in terms of what they are measuring. But although it may be obvious that a criterion-referenced test should be inter-

> nally consistent, it is not obvious how to assess the internal consistency. The classical procedures are not appropriate. This is true because they are dependent upon score variability. A criterion-referenced test should not be faulted if, when administered after instruction, everyone obtained a perfect score. Yet, that would lead to a zero internal consistency estimate, something measurement books don't recommend.
>
> In fact, even stranger things can happen in practice. It is possible for a criterion-referenced test to have a *negative* internal consistency index --and still be a *good* test. ...
>
> Other aspects of reliability are equally cloudy. Stability might certainly be important for a criterion-referenced test, but in that case, a test-retest correlation coefficient, dependent as it is on variability, is not necessarily the way to assess it. Some kind of confidence interval around the individual score is perhaps a partial solution to this problem.
>
> The reader should not misinterpret the above statements. If a criterion-referenced test has a high average inter-item correlation, this is fine. If the test has a high test-retest correlation, that is also fine. The point is *not* that these indices cannot be used to support the consistency of the test. The point is that a criterion-referenced test could be highly consistent, either internally or temporarily, and yet indices dependent upon variability might not reflect that consistency. (pp. 5-6)"

Clearly, the major issue that Popham and Husek consider is the very real possibility that a set of criterion-referenced test scores may not display much variance. In this case, the classical measures of reliability are apt to be inappropriate.. This is perhaps the most frequently cited reason for the need to develop new measures of reliability for criterion-referenced tests.

Another frequently cited reason for developing new indices is that criterion-referenced tests frequently employ a mastery cutting score that is intended to be independent of the distribution of observed (and true) scores. The presence of this cutting score argues that an important issue in the reliability of mastery tests involves the extent to which the test is a dependable instrument for assessing whether or not persons surpass the mastery cutting score.

These are the two most frequently cited reasons for pursuing the development of new measures of reliability for criterion-referenced tests. Our discussion of particular indices in the next section will build upon and, in some cases, further refine these reasons.

Criterion-Referenced Reliability Indices

The literature contains a number of suggested statistics for estimating the reliability of criterion-referenced test scores. In thissection we describe most of these indices and, when appropriate, we comment on their characteristics, strengths, and weaknesses.

Ivens' agreement indices. Ivens (1970) argues that measures of reliability for criterion-referenced tests should be independent of test score variance; therefore, the measures he proposes are based upon a consideration of different kinds of agreement.

First, Ivens considers reliability using the concept of within subject equivalence of total scores. "For each subject, the raw score for the two administrations, either test-retest or parallel forms, (is) converted into percent-correct scores. For each examinee, the absolute difference between the percent correct on the two administrations (is) obtained. ... The actual reliability index (consists) of reporting ... the percent of subjects with percent-difference scores of a given size or less (Ivens, 1970, p. 11)." This measure can be expressed algebraically as follows: Let

X_{ijl} = the response of person i (i = 1,2, ..., N),
to item j (j = 1,2, ..., K)
on test l (l = 1,2) where
l = 1 means the first administration of the test (or the first of the two parallel tests) and
l = 2 means the second administration of the test (or the second of the two parallel tests.

$$A_i = \begin{cases} 1 \text{ if } |X_{i.1} - X_{i.2}| <= c \\ 0 \text{ if } |X_{i.1} - X_{i.2}| > c \end{cases}$$

where c is some tolerance limit in the range 0 <= c <= 1.0.

Now, the reliability (in the sense of agreement) over persons, given a tolerance limit of c is:

$$AP(c) = (1/N) \sum_i A_i .$$

Note that AP(c) is a proportion, and there are as many possibly different values of AP(c) as there are values of c. If we plotted AP(c) against c, then we would observe that AP(c) is a monotonically non-decreasing function of c. Thus, in order to report AP(c) in its entirety, we should report something like a plot of AP(c) for 0 <= c <= 1.0. If this procedure is not followed, then, at a minimum, one could report several selected values of AP(c).

Second, Ivens considers test reliability as the average of the individual item reliabilities where item reliability is expressed by calculating the proportion of subjects whose item scores (pass-fail or correct-wrong) are the same on the test and the retest, or on the test and the parallel form. Using this line of reasoning the reliability for item j is defined as:

$$AI_j = (1/N) \sum_i A_{ij} \text{ , where}$$

$$A_{ij} = \begin{matrix} 1 \text{ if } X_{ij1} = X_{ij2} \\ 0 \text{ if } X_{ij1} \neq X_{ij2} \end{matrix}$$

Thus, test reliability is defined as:

$$AI = (1/K) \sum_j AI_j$$

$$= [1/(NK)] \sum_i \sum_j A_{ij}$$

$$= (1/N) \sum_i [(1/K) \sum_j A_{ij}]$$

The measure AI is very appealing in that it is a linear function of item reliabilities. Thus, for example, if one knows the reliability of each of the items in an item bank, then one can estimate the reliability for any test (i.e., for any subset of items that might be selected from the item bank).

Ivens' measures have several appealing characteristics. First, they are distribution-free. Second, they do not depend upon test score variance. Third, they are simple to calculate. Fourth, they can be used to calculate measures of stability or equivalence. Fifth, they are relatively easy to interpret.

From a different point of view, we note that these measures have certain characteristics that some may consider undesirable. First, they are not interpretable in terms of the ratio of true score variance to observed score variance; therefore, they are not measures of reliability under the classical test theory model. Second, Ivens' indices are not likely to provide a great deal of help to the researcher interested in estimating a person's true score, which is, indirectly, a typical function of a reliability index under the classical test theory model.

Berger-Carver mastery agreement. Berger (1970) and Carver (1970) consider a method similar to Ivens' agreement indices for assessing the reliability of a criterion-referenced test. In the Berger-Carver case, however, a subject's score is treated as a dichotomous variable; i.e., a subject is placed into a mastery or a non-mastery group depending upon whether the subject surpassed or failed to surpass some minimum performance level, or mastery score. On a test-retest or parallel forms basis, a subject's two scores constitute an agreement if they result in the same classification; and the reliability measure is the agreement proportion over subjects. Letting

A = the number of subjects who scored above the mastery cut-off on both the test and retest (or both parallel forms),

B = the number of subjects who scored below the mastery cut-off on both the test and the retest (or both parallel forms), and

N = the total number of subjects,

the Berger-Carver mastery agreement (reliability) measure can be expressed as:

BC-MA = (A + B)/N

A major conceptual difference between the BC-MA statistic and Ivens' indices is that the BC-MA statistic involves a mastery cut-off, while Ivens' indices do not.

Another difference is that, for the BC-MA statistic, a student's total test score functions as an intermediate score -- intermediate to scoring the student 1 (master) or 0 (non-master). In most other respects, the advantages and disadvantages noted for Ivens' measures apply to the BC-MA statistic as well.

Marshall's index of separation. Marshall (1973) proposed an index based upon the assumption that the population taking a criterion-referenced test is the union of two subpopulations, either of which may be empty. For the "knowledgeable" subpopulation, the expected value of a person's score is assumed to be equal to the number of items in the test; for the "not knowledgeable" subpopulation, the expected value of a person's score is assumed to be equal to zero. Marshall's index of separation is defined as:

$$SEP = 1.0 - (4/nN) \sum_{i=1}^{N} (X_i - X_i^2/n), \text{ where}$$

n = the number of test items,
N = the number of persons, and
X_i = a person's total score (number of items correct) on the test.

This index has a range of zero to one, and it is related to the variance of the total scores by the formula:

$$SEP = 1.0 - 4[\overline{pq} - ((N-1)/(n^2N))s_X^2], \text{ where}$$

$\overline{p}$ = mean proportion correct over items (i.e., mean item difficulty) and
$\overline{q} = 1 - \overline{p}$

Marshall notes that the index of separation stays constant at 1.0 when (a) total scores are all zero, (b) total scores are all n, and (c) total scores are half zero and half n. Thus, Marshall's index of separation obviates one of the objections to classical reliability indices for criterion-referenced tests -- namely, the classical formulas give a reliability of zero when the variance of total scores is zero.

Harris' index of efficiency. Harris (1972a) proposed an index of efficiency defined as:

$$\text{Eff} \doteq \frac{SS_b}{SS_b + SS_w} ,$$

where the between-groups and the within-groups sums of squares are determined by the two groups resulting from dichotomizing subjects into masters and non-masters. Harris states that the purpose of his index is "to measure how well the test sorts defined samples of students into (mastery and non-mastery) categories and possibly to measure its efficiency in this sense (Harris, 1972a, p. 4)." Harris points out that EFF can be viewed as the ratio of true score variance to observed score variance if a subject's true score is defined as the mean of that subject's group (mastery group or non-mastery group).

Hambleton-Novick indices. Hambleton and Novick (1973) state that "in most cases, the pertinent question (in criterion-referenced testing is whether or not the individual examinee has attained some specified degree of competence on an instructional performance task (p. 160)." Hambleton and Novick interpret the "specified degree of competence" as a mastery score.

In order to consider the reliability index they propose, we must first review their decision-theoretic approach to criterion-referenced measurement. This approach is considerably different from other approaches reported in the literature, and, in the opinion of this author, the decision-theoretic approach has much to reccommend it, at least from a theoretical viewpoint for mastery testing. Their approach is similar, in some respects, to the "quota-free" selection problem discussed in Cronbach and Gleser (1965). "That is, there is no quota on the number of individuals who can exceed the cut-off scores or threshold on a citerion-referenced test (Hambleton and Novick, 1973, p. 163)." Again, it should be noted that Hambleton and Novick are using the term "criterion-referenced test" in the sense of "mastery test."

Quoting from Hambleton and Novick (1973):

> "The primary problem in the new instructional models, such as individually presecribed instruction, is the one of determining if π_i, the student's true mastery level, is greater than a specified standard π_o. Here, π_i is the "true" score for an individual i in some particular well specified content domain. It *may* represent the proportion of

items in the domain he could answer successfully. Since we cannot administer all items in the domain, we sample some small number to obtain an estimate of π_i, represented as $\hat{\pi}_i$. The value of π_o is the somewhat arbitrary threshold score used to divide individuals into the two categories described earlier, i.e., Masters and Non-masters.

Basically then, the examiner's problem is to locate each examinee in the correct category. There are two kinds of errors that occur in this classification problem: False positives and false negatives. A false-positive error occurs when the examiner estimates an examinee's ability to be above the cutting score when, in fact, it is not. A false-negative error occurs when the examiner estimates an examinee's ability to be below the cutting score when the reverse is true. The seriousness of making a false-positive error depends to some extent on the structure of the instructional objectives. It would seem that this kind of error has the most serious effect on program efficiency when the instructional objectives are hierarchial in nature. On the other hand, the seriousness of making a false-negative error would seem to depend on the length of time a student would be assigned to a remedial program because of his low test performance. (Other factors would be the cost of materials, teacher time, facilities, etc.) The minimization of expected loss would then depend, in the usual way, on the specified losses and the probabilities of incorrect classification. This is then a straightforward exercise in the minimization of what we would call threshold loss.

In an attempt to view the above discussion in a more formal manner, suppose we take some criterion level π_o, and define a parameter ω such that

$$\omega = 1 \text{ if } \pi \geq \pi_o$$

$$\omega = 0 \text{ if } \pi < \pi_o \quad .$$

Persons having ω values of one are those who have true ability levels equal to or greater than the criterion level π_o, and those having ω values of zero are those whose π values are below π_o. Now if we obtain an estimate of π_i, then an estimate of ω would be obtained in the following way:

$\hat{\Omega} = 1$, if $\hat{\pi}_i \geq \pi_o$ and

$\hat{\Omega} = 0$, if $\hat{\pi}_i < \pi_o$.

> Defining our error of estimation as $(\hat{\Omega} - \omega)$, the difference between the estimated and the true value, it is clear that the error takes on one of three values; +1, -1, 0, corresponding to whether we make a false-positive error, a false-negative error, or a correct classification. Also, note that the squares of the errors and their absolute values are identical. Thus, any procedure that minimizes squared-error loss (SEL) in the ω metric also minimizes absolute-error loss (AEL) in that metric. The criterion-referenced measurement problem is, thus, one of determining an estimator $\hat{\Omega}$ of ω by determining an estimator $\hat{\pi}$ of π with a threshold loss function and converting this to an estimate of ω . ... Note that with threshold loss, the estimate $\hat{\pi}$ of π is not a single number but one of two intervals $[0,\pi_o)$ or $[\pi_o, 1]$. ... The minimization of SEL and AEL in the ω metric is equivalent to the minimization of threshold loss for π in the special case where the losses associated with false positives and false negatives are equal (pp. 163-164)."

In order to make use of the procedure indicated above, one must obtain estimates for the π_i . In order to accomplish this, Hambleton and Novick suggest a Bayesian solution that involves using the "direct information provided by the student's ... score (and) the collateral information contained in the test data of other students. (Another possibility and one worthy of future research is that of using the student's other subscale scores and provious history as collateral information.) (p.165)."

Using the above approach, Hambleton and Novick suggest two reliability coefficients. First, assuming the existence of two tests that are parallel in the ω metric, let

$\hat{\Omega}_{1i}$ = score for person i on first parallel test and

$\hat{\Omega}_{2i}$ = score for person i on second parallel test.

Then, one possible reliability coefficient is the correlation, over persons, for the two parallel tests; i.e.,

HN-CORR = $\text{corr}(\hat{\Omega}_{1i}, \hat{\Omega}_{2i})$.

Another measure of reliability they suggest is the proportion of times that the same decision is made with the two parallel measurements; i.e.,

$$\text{HN-MA} = A/N \text{ , where}$$

A = number of times $\Omega_{1i} = \Omega_{2i}$, and
N = total number of subjects.

Clearly, these two measures are measures of equivalence; analogous measures of stability can be constructed directly.

The critical problem in the Hambleton and Novick procedure involves the estimation of the π_i scores, which are the "true" scores for the individuals.[1] ("True score" is never defined by Hambleton and Novick; therefore, we assume here that they mean true score in the classical sense.) It should be noted that the authors' suggested Bayesian solution to the problem is, at the present time, an unsolved problem, since the most appropriate available procedure (Novick, Lewis, and Jackson, 1973) does not use a threshold loss function, according to Hambleton and Novick.

Livingston's coefficient. Livingston's (1972b) coefficient has undoubtedly received more attention (and criticism) than any other criterion-referenced reliability measure that has been reported in the literature. See, for example, Livingston (1972a,b,c), Harris (1972b, 1973), and Shavelson et al. (1972). Livingston's reliability coefficient can be expressed as:

$$\text{LIV} = \frac{r_{tt}V(X) + (\bar{X} - C)^2}{V(X) + (\bar{X} - C)^2} \text{ , where}$$

r_{tt} = any "norm-referenced reliability coefficient" based upon the classical test theory model,
C = a mastery cutting score,
$\bar{X}$ = mean score over persons, and
$V(X)$ = variance over persons.

Much of the discussion of LIV has involved some degree of misunderstanding about the nature of the coefficient; therefore, let us list a few characteristics of LIV:

(a) LIV involves a consideration of the expected squared deviation of a person's score from C, as distinct from the expected squared deviation of a person's score from $\bar{X}$ (the latter being a definition of variance). Therefore, LIV involves an "atypical" squared error loss function.

(b) The range of LIV is [0,1], and LIV is, therefore, similar to classical reliability coefficients, r_{tt}, in this respect. LIV equals r_{tt} when $C = \bar{X}$, and LIV > r_{tt} when $C \neq \bar{X}$.

(c) LIV is identical to a classical reliability coefficient when that coefficient is based upon two populations with means equally distant above and below C (Harris, 1972b).

(d) The classical standard error of measurement is the same for both LIV and r_{tt}; therefore, the (usually) larger value of LIV does not imply a more dependable estimate of a person's true score, in the classical sense, nor does it imply a more dependable determination of whether or not a true score falls above or below C (Harris, 1972b). However, the usually larger value of LIV does imply "a more dependable overall determination of whether each true score falls above or below the criterion level, when this decision is to be made for every individual score in the distribution (Livingston, 1972a, p. 31)."

(e) In general, there is no algebraic transformation of the observed test scores that produces a set of scores such that, when these scores are used in a classical reliability formula, the result equals LIV. One is tempted to think that this might be true if one used the deviation scores $X_i - C$, but, since these scores are linear transformation of the X_i scores, the classical reliability of the deviations scores equals the classical reliability of the original X_i scores.

In the opinion of this author, the net result of these observations seems to be that LIV has some useful descriptive properties, _if_ one accepts that Livingston's "atypical" squared error loss function is meaningful and appropriate. (Hambleton and Novick, 1973, are two researchers who seriously question the kind of squared error loss used by Livingston.) However, it is clear that LIV relies upon test score variance, and this characteristic is a negative factor, in the minds of many researchers. Also, it is clear that LIV does not enhance our ability to estimate a person's true score, even when LIV is very much greater that its corresponding classical reliability coefficient.

In short, this author sees no compelling reason for generally abandoning the use of LIV as some might suggest; however, this author also feels that LIV should not be considered as the answer to the question of measuring the reliability of a criterion-referenced test. Also, users of LIV should be very careful to interpret and use this coefficient correctly. It is extremely easy to increase the value of LIV by moving C farther away from the mean; however, this should be done if and only if there is a substantively defensible reason for doing so. Finally, the reader should note that LIV is really a coefficient for mastery testing, not for criterion-referenced testing, in general.

Ozenne's sensitivity indices. Ozenne (1971) claims that in a criterion-referenced testing situation the important question is, "How effective has instruction been?" The rationale for his first sensitivity index lies in "the implicit assumption that if there is a difference in level of response on the two (testing) occasions, ... such a difference is due to the intervening instruction (Ozenne, 1971, p. 17)." In Ozenne's model the two testing occasions under consideration are the pretest and the posttest. More explicitly, the model under consideration is:

$$Y_{ijk} = \pi + \alpha_j + \beta_k + (\alpha\beta)_{jk} + e_{ijk}, \text{ where}$$

π = population parameter,

α_j = effect due to persons, $j = 1,2, \ldots, N$;

β_k = effect due to occasions (i.e., effect due to instruction), $k = 1,2$;

$(\alpha\beta)_{jk}$ = effect due to interaction of examinees (persons) and occasions factors; and

e_{ijk} = error of measurement.

Using this model, Ozenne's first sensitivity index is:

$$\text{SENS}_1 = \frac{MS_{\text{occasions}} - MS_{\text{interaction}}}{MS_{\text{occas.}} - MS_{\text{inter.}} + N \cdot MS_{\text{error}}}$$

This index is, in effect, the variance due to instructional effects (the occasions effect) divided by the sum of the variances due to instructional effects and errors of measurement.

Ozenne's second index of sensitivity is given by the formula:

$$\mathrm{SENS}_2 = \frac{\mathrm{MS}_{\mathrm{treatment}} - \mathrm{MS}_{\mathrm{subjects\ w.\ treatment}}}{\mathrm{MS}_{\mathrm{treat.}} - \mathrm{MS}_{\mathrm{subjs.w.treat.}} + N'\mathrm{MS}_{\mathrm{error}}}$$

This index is intended to be used when one has two different treatment groups -- one group receiving instruction and the other not. The underlying statistical model is:

$Y_{ijk} = \pi + \beta_k + \alpha_{j(k)} + e_{ijk}$, where

π = population parameter;

β_k = effect due to treatments, k = 1,2;

$\alpha_{j(k)}$ = effect due to persons nested within treatment k; and

e_{ijk} = error of measurement.

Other suggested indices. In addition to the Berger-Carver mastery agreeme : statistic, Carver (1970) suggests that, "the reliability of a single form of a criterion-referenced device could be estimated by administering it to two comparable groups. The percentage that met the criterion is one group could be compared to the percentage that met the criterion in the other group (p. 56)."

Cox and Graham (1966) and Ferguson (1971) suggest use of the coefficient of reproducibility for reliability estimation when the criterion-referenced test items are assumed to form a Guttman Scale.

Discussion. It seems appropriate to suggest some statements, of a comparative nature, concerning the above indices.

First, all of the above indices, except those suggested by Ivens, Marshall, and Ozenne, are, more precisely, indices for mastery tests, since these indices depend, one way or the other, on the specification of a mastery cutting score. For these mastery test reliability indices, it is important to observe that the fundamental or primary student score under consideration is often ambiguous. For example, is the fundamental score the number (or percentage) of items correct, or the extent to which this score is above or below the mastery score, or merely whether or not this score is above or below the mastery score? Another way to view this issue is to ask the question, "What is the appropriate error of measurement?" Only Hambleton and Novick (1973) address this issue in any depth. In short, there is a considerable lack of test theoretic justification (classical or otherwise) for many of the suggested reliability indices for mastery tests and, for that matter, criterion-referenced tests, in general.

Second, several of the above indices (Marshall's, Harris', Livingston's, and possibly Hambleton and Novick's) depend, directly or indirectly, upon the variance of student scores. Many researchers feel that the variance of student scores should exert no, or minimal, influence upon judgments about a criterion-referenced test's reliability or validity.

Third, since Marshall's, Harris', and Livingston's indices involve only one administration of a test, they cannot be considered measures of stability. For the most part, these indices seem to be measures of the extent to which the test is dependable in its ability to classify subjects as masters or non-masters. Therefore, in a sense, these measures are analogous to, what are usually called measures of internal consistency. Also, at least Livingston's index can be interpreted as a measure of equivalence.

Fourth, only Ozenne's SENS$_1$ index incorporates both pre- and posttest scores; therefore, this index may appear to have the potential for assessing the reliability of change scores, whereas the other indices clearly do not have this potential. However, Ozenne's SENS$_1$ index is primarily a measure of instructional effectiveness, not a measure of the reliability of criterion-referenced change scores. Also, it should be noted that Ozenne's second index is not really a reliability index either; Ozenne's second index is merely another measure of instructional effectiveness. One could certainly argue, therefore, that Ozenne's indices should not even be discussed in this chapter; eventhough other researchers have mistakenly considered Ozenne's indices to be measures of reliability.

Marshall (1973) provides additional information and insight into the characteristics and function of many of the above indices.

In the opinion of this author: (a) Hambleton and Novick's indices have the most appealing theoretical rationale of those indices proposed for mastery tests, but one important statistical problem remains to be solved before these indices will be generally useful, (b) Livingston's index is mathematically similar to classical reliability indices, but it employs a questionable theoretical basis and is somewhat difficult to interpret, (c) the indices attributable to Harris and Marshall may have practical utility, but this has not yet been demonstrated, and, in addition, both of these indices may be questionable from a theoretical point of view, (d) Ozenne's indices are not really reliability indices, eventhough they have been treated as such by some authors, and (e) Iven's indices, as well as the Berger-Carver index, are appealing in several respects, and Iven's indices are th? most appropriate available indices for a criterion referenced test when a mastery cutting score is not employed, but none of these indices has yet received sufficient critical examination by researchers and practitioners. In short, many important issues surrounding the reliability of criterion-referenced measures remain unsolved problems, or, at best, these issues have not yet received adequately complete treatment in the literature.

CHAPTER V

Criterion-Referenced Item Analysis and Revision Procedures Employing Classical Scoring

The differences between criterion-referenced and norm-referenced testing have led most researchers to conclude that norm-referenced item analysis procedures are of questionable value in criterion-referenced testing situations. (See, for example, Popham and Husek, 1969, Popham, 1971, Cox and Vargas, 1966, and Brennan, 1970.)

Yet, clearly, a criterion-referenced test can be no better than the items it contains. Therefore, if we are to develop reliable and valid criterion-referenced tests, we need statistics to describe the performance of students on items, we need statistics for assessing item reliability and validity, and we need procedures for identifying poor or undesirable criterion-referenced test items. These topics are the subject of this chapter. Specifically, in this chapter we will consider: (a) item statistics for criterion-referenced tests, (b) a procedure for identifying criterion-referenced test items and instruction that require revision, and (c) the use of item analysis tables in criterion-referenced testing situations. In practically all cases, in this chapter, the statistics and procedures we discuss entail the use of the classical correct/wrong scoring procedure. Other scoring procedures are considered in Chapter VI.

The subject of item analysis and revision procedures for criterion-referenced tests is especially crucial and especially difficult. It is especially crucial in that the validity of a criterion-referenced test is very closely tied to the validity of the individual items. It is especially difficult in that: (a) there are few, if any, objective, empirically-based criteria for "good" criterion-referenced test items, and (b) even if such criteria did exist, empirical data can identify items that may require revision, but empirical data can seldom, if ever, dictate that an item must be revised or eliminated. Thus, at least at the present time, any total evaluation of a criterion-referenced test item necessitates a considerable amount of subjective judgment on the part of subject matter specialists.

The statistics and procedures discussed below have been culled from the literature or developed by the author. Thus, they represent a statement of the state-of-the-art in criterion-referenced item analysis and revision procedures, basically from an empirical point of view. However, it should be understood that there is considerable discussion and even some disagreement among researchers concerning the applicability of these statistics and procedures. Much work remains to be done.

Item Statistics

In this section we consider item statistics relevant to criterion-referenced testing. Most of the statistics discussed here are reported in the literature; the others were developed by the author and are offered for consideration. The reader will note that we consider two kinds of statistics for items: (a) measures of state (i.e., measures that reflect student performance at one point in time) and (b) measures of change (i.e., measures that reflect student performance at two points in time). Also, for both of these possibilities we consider statistics for describing the reliability and validity of an item.

Most of the literature that discusses criterion-referenced item statistics treats these statistics as measures of state; however, criterion-referenced tests are often used to assess change, especially as the issue of change relates to the effectiveness of an instructional system (see Chapter I).

Measures of state. For the most part, in criterion-referenced testing, measures of state are expressed as difficulty levels or, less frequently, as error rates. The difficulty of an item is defined as the proportion of students who get an item correct. As such, the term "difficulty level" is somewhat of a misnomer in that if difficulty level is high then the item is easy, and if difficulty level is low then the item is "difficult." Since difficulty level is a proportion, its range is zero to one.

Error rate is defined as the proportion of students who get an item incorrect; it is mathematically equal to one minus the difficulty level, and its range is also zero to one. Thus, error rate contains all of the information that difficulty level contains, and error rates do not suffer from the interpretation problem encountered with difficulty levels.

Measures of change. It is not our intent here to indulge in a lengthy discussion of the measurement of change. We have previously discussed this issue to some extent, and it will be a subject of further discussion later. Here we merely want to identify major references relating to the measurement of change in criterion-referenced testing situations.

One of the earliest empirical studies using criterion-referenced test data was performed by Cox and Vargas (1966). The index they considered was simply the difference between posttest difficulty level and pretest difficulty level. Hambleton and Gorth (1971) and Popham (1971) have also examined this index, and, in addition, Popham (1971) has considered various other statistics that depend upon change scores. For the most part, these authors have treated the indices they analyzed in a manner similar to the way discrimination indices are treated in norm-referenced testing. That is, the indices have been viewed primarily as statistics for identifying "bad" or "atypical" criterion-referenced test items.

It should be pointed out that one could also argue that these indices are measures of instructional effectiveness. In fact, in criterion-referenced testing situations in instructional environments, Ivens (1970) and Brennan (1970) treat measures of change primarily as measures of instructional effectiveness, and only secondarily as indices for identifying poor criterion-referenced test items. The measure of item change proposed by Ivens (1970) has been introduced in Chapter IV and will be discussed again below. Brennan (1970) has suggested the consideration of indices called "percentage of maximum possible gain" and "percentage of maximum possible effectiveness."

Item reliability -- measures of state. For the classical test theory model, the reliability of an item (in an internal consistency or equivalence sense) is usually calculated by determining the intraclass correlation coefficient using Hoyt's (1941) analysis of variance framework. This technique has been considerably extended recently by the work of Cronbach et al (1972). The intraclass correlation coefficient may be an appropriate measure for criterion-referenced item reliability (in an equivalence sense) if: (a) the variance of the total scores over items is not close to zero, (b) all items are measures of the same objective, and (c) one accepts (and the data fulfil) the implicit assumptions entailed in using the intraclass correlation coefficient as a measure of criterion-referenced item reliability.

In the following paragraphs we consider a number of indices that have been proposed specifically for the purpose of calculating item reliability in criterion-referenced situations.

Ivens' (1970) measure denoted AI_j in Chapter IV provides us with a measure of item reliability, in either an equivalence or stability sense, when all subjects take both parallel items or when all subjects are administered the same item twice, respectively. Since this index is a reliability (R) index for a measure of state (S), let us denote this index as RS.

Now suppose we have two items which are intended to be equivalent measures of a particular objective. It is not always feasible or desirable to have subjects take both items, yet we usually do want a measure of item reliability. Let us now consider a procedure, offered by this author, for obtaining item reliability, given two supposedly parallel items, when (a) all students are randomly assigned to one of two groups, and (b) students in group one respond to the "first" item and students in group two respond to the "second" item.

Let us denote persons in group one as:

$$A_j \ , \ j = 1,2, \ldots, n$$

and persons in group two as:

$$B_k \ , \ k = 1,2, \ldots, n \ .$$

Now, if we considered students A_j and B_k to be the same persons when $j = k$, we could calculate the index RS and have a measure of item equivalence. However, since the order of the persons' subscript is arbitrary, the resulting index is only one of a large number of possibilities; for example, we could just as well have considered A_1 and B_2, A_2 and B_3, ..., A_n and B_1 to be the same person. However, we can extend this rationale to obtain what may be a reasonable index.

The procedure is as follows: (a) calculate RS for each distinct way of pairing persons in group one with persons in group two and (b) average the RS indices in order to obtain RS*, which will denote the desired measure of equivalence. This procedure can also be applied when there are unequal numbers of subjects in groups one and two, say

n_1 = number of subjects in group one, and

n_2 = number of subjects in group two, where

$n_2 \geq n_1$.

The procedure indicated above is, however, cumbersome because there are

$$(n_2)(n_2 - 1)(n_2 - 2) \cdots (n_2 - n_1 + 1)$$

different ways of pairing the n_1 subjects in group one with subjects in group two.

Therefore, it is fortunate that the above procedure is mathematically equivalent to calculating RS when we treat

$$\begin{array}{cccc}
A_1 \text{ and } B_1 & A_2 \text{ and } B_1 & \cdots & A_{n_1} \text{ and } B_1 \\
A_1 \text{ and } B_2 & A_2 \text{ and } B_2 & \cdots & A_{n_1} \text{ and } B_2 \\
\vdots & \vdots & \cdots & \vdots \\
A_1 \text{ and } B_{n_2} & A_2 \text{ and } B_{n_2} & \cdots & A_{n_1} \text{ and } B_{n_2}
\end{array}$$

as $n_1 n_2$ different subjects. That is, we examine all possible pairs of subjects, where a pair is defined as one person from each group. If the item score for both persons in a pair is the same, then that constitutes an agreement, and RS* is the number of agreements divided by the total number of pairs. Letting

c_1 = number of subjects in group one who got "first" parallel item correct, and

c_2 = number of subjects in group two who got "second" parallel item correct,

it can be shown that:

$$RS^* = \frac{c_1c_2 + (n_1 - c_1)(n_2 - c_2)}{n_1n_2}$$

$$= p_1p_2 + q_1q_2$$

$$= 1 - p_1 - p_2 + 2p_1p_2 \text{ , where } \begin{array}{l} p_1=c_1/n_1 \\ p_2=c_2/n_2 \\ q_1=(n_1-c_1)/n_1 \\ q_2=(n_2-c_2)/n_2 \end{array}$$

The index RS* has a range of zero to one.

Another measure of item reliability (strictly in the sense of equivalence) is suggested by Sabers and Kania (1972). Let us identify the two supposedly parallel items as items j and j', and let us display the data in the following form:

		Item j'	
		Pass	Fail
Item j	Pass	A	B
	Fail	C	D

where A = number of students who passed both items,
B = number of invalid passes on item j,
C = number of invalid passes on item j', and
D = number of students who failed both items.

Sabers and Kania define the "index of item precision" for items j and j', respectively, as:

$$P_j = 1 - B/N$$

and $$P_{j'} = 1 - C/N \text{ ,}$$

where N is the total number of subjects.

Using these two indices of item precision, Sabers and Kania define the item reliability coefficient as:

$$RS^{**} = 0.5(P_j - P_{j'})(1 - |P_j - P_{j'}|) .$$

This coefficient (which they call the KI coefficient of item equivalence) has a range of zero to one. Sabers and Kania claim that the higher the value of RS** "the greater the degree of agreement between the decisions made by the two forms."

The reader will note that all of the above techniques are applicable only when we have two parallel items or two administrations of the same item. A procedure suggested by Brennan and Stolurow (1971) is applicable for any number of parallel items. Brennan and Stolurow note that, in classical test theory, the statistical criteria for K parallel tests are that the K means, the K variances, and the K(K-1)/2 intercorrelations be equal. When one has K items, instead of K tests, the same criteria would seem to be appropriate. If item scores constitute a multivariate normal distribution, then the above assumptions can be tested using a procedure developed by Wilks (1946). However, in most criterion-referenced testing situations one cannot justifiably assume that item scores constitute a multivariate normal distribution. In such cases, in order to test the equality of the K means and the K variances one can use Cochran's Q test (see Siegel, 1956); however, this author knows of no appropriate statistical procedure for testing the equality of the intercorrelations in the absense of a multivariate normal distribution of item scores. Therefore, in most criterion-referenced testing situations, at least at the present time, researchers will have to make subjective judgments about the equality of item intercorrelations.

Item validity -- measures of state. The usual measure of item validity is a discrimination index, which compares item scores with scores on some criterion. For most criterion-referenced tests, total test score is usually the only available, appropriate criterion. A number of correlational type discrimination indices have been reported in the literature; however, for criterion-referenced testing they have the disadvantage of being severely affected by small amounts (or lack of) variance in the distribution of item and/or test scores. (See Brennan, 1970, for a more complete discussion of such indices.) Therefore, this author recommends use of the following index discussed in detail by Brennan (1972):

$VS = (c_u/n_u) - (c_l/n_l)$, where

c_u = the number of students in the upper group who got the item correct,

c_l = the number of students in the lower group who got the item correct,

n_u = the total number of students in the upper group, and

n_l = the total number of students in the lower group.

In Brennan (1972) the VS index is called the B index; here, we have chosen the designation "VS" to indicate that the index relates to item validity (V) for measures of state (S). This index has a lower limit of -1 and an upper limit of +1. For mastery testing, upper and lower groups would usually be defined in terms of the mastery cutting score. It may be appropriate, in some cases, to eliminate from consideration students close to the mastery cutting score, since such students are on the borderline of mastery.

Both Popham and Husek (1969) and Brennan (1972) agree that for criterion-referenced testing: (a) negatively discriminating items are undesirable, (b) nondiscriminating items are not necessarily bad items, and (c) positively discriminating items may indicate ineffective instruction. Brennan (1972) also points out that if all students get an item correct, then the VS index equals zero. Therefore, if it is desirable that all students get an item correct, the the "ideal" value of VS is zero, and, hence, the ideal item is a nondiscriminating item. Following this line of reasoning, even positively discriminating items (and certainly negatively discriminating items) indicate that either the test item or instruction may require revision.

Item reliability -- measures of change. In order to address this issue, let us define the following:

X_{i1} = pretest response (0,1) of person i to the first administration of an item (or to the first of two parallel items),

X_{i2} = pretest response (0,1) of person i to the second admininstration of an item (or to the second of two parallel items),

Y_{i1} = posttest response (0,1) of person i to the first administration of an item (or to the first of two parallel items),

Y_{i2} = posttest response (0,1) of person i to the second administration of an item (or to the second of two parallel items),

$D_{i1} = Y_{i1} - X_{i1} = 0, 1,$ or -1, and

$D_{i2} = Y_{i2} - X_{i2} = 0, 1,$ or -1 .

Now, the reliability of an item as a measure of change can be expressed as the number of subjects for whom $D_{i1} = D_{i2}$ divided by the total number of subjects. To be consistent with our designation for other indices, let us denote this index of item reliability (R) for a measure of change (C) as RC; the reader will note that the range of RC is from zero to one.

It is important to notice that , for the RC index, if

$$D_{i1} = D_{i2} = 0 \text{ for all subjects } i,$$

then RC = 1; i.e., the change score reliability of the item is perfect eventhough, for every student, no change has occurred. This is not a contradiction. The fact that RC = 1 merely indicates that the item is perfectly reliable when used as a measure of change; this fact does not say anything about the amount of change or the direction of change.

Item validity -- measure of change. In order to construct such an index, we must have some criterion for change. One possible criterion (although not necessarily a good one) is the set of student scores, each of which is an average item change score, where an item change score is defined as posttest item score minus pretest item score. Using the notation in the previous section and replacing the second subscript by an item subscript j, a person's average change score is:

$$D_{i\cdot} = (1/K) \sum_{j=1}^{K} (Y_{ij} - X_{ij}) \text{ , where}$$

$j = 1, 2, \ldots, K$ items, and

$-1 \leq D_{i.} \leq 1$.

Using the above scores (or some other change score criterion if available and appropriate) one can define upper and lower groups. Then one can compare the trichotomized item change scores with the dichotomized criterion change scores using the following table:

	Item Change Score		
	-1	0	1
Upper Group	P(U,-1)	P(U,0)	P(U,1)
Lower Group	P(L,-1)	P(L,0)	P(L,1)

In this table P(U,-1) means the proportion of students in the upper group who got an item change score of -1; the other cells are interpreted in a similar manner. Using the above table one can examine the validity of the item as a measure of change; however, no single statistic with a range of -1 to +1 appears to be readily available from this table for the purpose of assessing the extent to which the item is a valid measure of change. Of course, one could obtain a single statistic merely by correlating item and criterion change scores, but the appropriateness of such a procedure needs to be examined for criterion-referenced testing situations.

Other possible indicesfor assessing item validity as a measure of change include Jenkins' (1956) triserial correlation coefficient and Saupe's (1966) change index. The latter is defined as:

$\mathrm{Corr}(Y_{i+} - X_{i+}, Y_{ij} - X_{ij})$, where

Y_{ij} = posttest score for person i on item j,

X_{ij} = pretest score for person i on item j,

Y_{i+} = total number of items correct on the posttest for person i, and

X_{i+} = total number of items correct on the pretest for person i .

Also, Ivens (1970) suggests two indices that might be used to assess item validity as a measure of change; however, Ivens' indices involve three sources of information -- pretest, posttest, and retest (or retention test).

It should be noted that the interpretation of any of the above indices is, or course, confounded by the presence of intervening instruction between pretest and posttest. Therefore, if the item does not appear to be valid when used to measure change, the problem may lie with the item, the instruction, or both.

A Decision Process for Identifying Criterion-Referenced Items and Instruction that Require Revision

In order to put the proposed decision process into a conceptual context, let us assume that we have an instructional program teaching a set of terminal objectives. Chronologically, each terminal objective is tested by a pretest item that occurs before the objective has been taught and a posttest item that occurs "some time after" the objective has been taught. Furthermore, we will assume that all of the items testing any objective are identical or equivalent.

In the final analysis, using item performance data, we want to identify those test items and sections of instruction (relevant to a given objective) that require revision. The decision process we propose will not necessarily tell the evaluator how to revise items and/or instruction, but the process will provide objective rules for deciding what to revise. (A previous version of the process proposed here is provided by Brennan and Stolurow, 1971).

Types of data and decision. Most of the decision rules discussed below make use of error rates and discrimination indices. An observed error rate for an item is the proportion of subjects who get the item incorrect; therefore, error rate is equal to one minus difficulty level. There are a number of discrimination indices that have been proposed in the literature; however, the applicability of many of them in criterion-referenced testing situations is open to question. Therefore, in general, we suggest using Brennan's (1972) B discrimination index (designated as VS in the previous section of this chapter).

For many of the proposed decision rules we will assume that error rates are classified as either high (H) or low (L), and that the evaluator predetermines an appropriate cut-off point between high and low error rate. For any given objective, the cut-offs for the error rates discussed below must be identical in order to apply the rules that will be specified. Also, in most cases, the cut-offs chosen will probably be the same for all objectives; however, occasions can arise when certain objectives should have a higher (or lower) error rate cut-off than other objectives. For example, items testing very crucial objectives might be assigned a cut-off of 0.10, while other items might have a cut-off of 0.25.

Discrimination indices will be classified as either positive (+), negative (-), or non-discriminating (0). By positive and negative indices we mean indices that discriminate significantly (at some appropriate α level) in the positive and negative directions, respectively.

Before instruction we can obtain two kinds of data for each objective that has a pretest item:

(a) the Theoretical Error Rate (TER), which is the expected proportion of students getting a pretest item incorrect simply on the basis of random guessing; i.e., if "a" is the number of possible answers to an item, then

$$\text{TER} = (a - 1)/a \ .$$

For example, if an item has five alternatives, we would expect 80 percent of the students to get the item incorrect simply by guessing randomly. Items that have a virtual infinitude of possible answers have TER = 1; however, the evaluator should be careful not to assume that every free-response or open ended test item has TER = 1. Very often such items are so worded that only two or three answers are possible, in which case TER = 0.50 or TER = 0.67.

(b) the Base Error Rate (BER), which is the observed proportion of students getting a pretest item incorrect.

After instruction we can obtain two types of data for each objective that has a posttest item: (a) the Posttest Error Rate (PER) and (b) the Posttest Discrimination Index (PDI).

In subsequent sections we will anlayze the decisions that can be made on the basis of the above data. Then we will discuss the decisions that can be made based upon the arithmetic differences between various error rates.

For each decision rule presented we will give our reasons for specifying whether test items or instruction relevant to a give objective should be revised (R), questioned (?), or not revised (NR). These decisions should not, however, be interpreted too strictly; the evaluator will still have to use some degree of subjective judgment. For example, when we say, in subsequent discussions, that an item should be revised (R), we mean that our best guess on the basis of the data is that the item should be revised, but the evaluator must make the final decision. Also, when we say that an item (or instruction) is questionable (?), we mean that the data are not sufficient to make a definite judgment about whether or not the item (or instruction) should be revised.

One additional consideration deserves mention. Ideally, one would validate his test items prior to using them in an instructional system; however, this is often not feasible, especially when criterion-referenced tests are used in an instructional system. Therefore, in most cases, evaluation must take into account the possible invalidity of both test items and instruction. For this reason, most of the decision rules that will be presented are based upon the assumption that we have no a priori reason to believe that test items are more valid than instruction or vice-versa.

Pretest data. It is not likely that only pretest data would be used to make decisions about test items, yet it is useful to consider the types of decisions that are appropriate on the basis of such data.

Rule 1: If TER and BER are both the same (i.e., H,H or L,L) then no necessity for revision is indicated. In this case, the observed error rate (BER), which is not affected by instruction, is approximately the same as the expected error rate (TER).

Rule 2: If TER is low (L) and BER is high (H), there is no indication that revision is required. This rather anomalous case could arise if the particular objective for the item involved concepts that are typically misunderstood. For example, many students (in the author's opinion) believe that "inflammable" and "flammable" have different meanings. If an item were constructed testing whether or not "inflammable" and "flammable" have the same meaning, and if this item were given prior to instruction, it is quite possible that more students would get the item incorrect than we would expect on the basis of the theoretical error rate (TER). In this case, there is no reason to revise the item; rather, we expect that the instruction will correct the students' misinformation.

Rule 3: If TER is high (H) and BER is low (L), then the item will probably need to be revised. In this case, students, without benefit of instruction, are performing considerably better than expected. It appears that the item itself may be teaching or that one or more distractors are so easy that many students can pick the correct answer largely by a process of elimination. It is also possible that the item is not at fault and the objective, while being easy for most of the students, is considered to be an integral part of the total set of objectives. In this case, of course, the item would not be revised.

These rules, as well as all other rules that will be discussed, are given in abbreviated form in Table 5-1.

Posttest data. As a result of administering a posttest two types of data can be collected: the Posttest Error Rate (PER) and Posttest Discrimination Index (PDI). Since these data are collected after instruction, theoretically decisions can be made about either test items or instruction or both. However, from a practical point of view, if revision seems to be required, if is difficult to specify with any confidence that the fault lies solely with the test item or solely with instruction. In short, based upon posttest data, we can usually say whether or not something is wrong, but given only the posttest data it is difficult to pinpoint the problem.

Rule 4: If PER = L and PDI = 0, then neither the item nor the instruction need to be revised. This is the best possible situation, since the optimal conditions for both error rate and discrimination index are fulfilled; i.e., at the end of instruction we hope that

TABLE 5-1

Rules for Decision-Making

Rule No.	Error Rates				Decision[a]	
	TER	BER	PER	PDI	Item	Instruction
1	H	H			NR	--
	L	L			NR	--
2	L	H			NR	--
3	H	L			R	--
4			L	0	NR	NR
5			L	+	?	?
			L	-	?	?
6			H	-	R	R
7			H	+	?	R
			H	0	?	R
8	DER > 0[b]				R	--
	DER ≤ 0[c]				NR	--
9		PMPG $< c$[d]			--	R
		PMPG $\geq c$[d]			--	NR

[a] "NR" means no revision required; "R" means revision is required; "?" means the data are not sufficient to make a sound judgment about whether or not revision is required.

[b] DER is significantly greater than zero for a one-tailed test of significance.

[c] DER is not significantly greater than zero for a one-tailed test of significance.

[d] c is a cut-off chosen by the evaluator.

most of the students get the posttest item correct (PER =L), and that the item is non-discriminating (PER = 0).

Rule 5: If PER = L and PDI = + or -, then both the item and the instruction are questionable. The fact that PDI is clearly non-zero indicates a possible need for revision.

Rule 6: If PER = H and PDI = -, then both the item and instruction should be revised, since PER = H and PDI = - is the worst possible situation that can occur. It is possible that either the item or the instruction is at fault, but not both; however, we assume here that the most universally applicable decision is to check both the item and the instruction to see what revisions are needed.

Rule 7: If PER = H and PDI = + or 0, then the instruction should be revised and the item should be questioned. Whenever error rate is high after instruction, something is wrong, but without additional information we do not know whether the fault definitely lies with the item or the instruction. However, the author believes that evaluators are apt to be more confident about test items than they are about instruction; it is also possible that the test items have been previously validated or partially validated. Therefore, in this case, it seems reasonable to place a less stringent decision on the item than on the instruction. It should be noted, however, that perceptions can be biased; i.e., the test item could be at fault. It is certainly advisable to analyze the nature of any validation or prevalidation activity for its applicability in the present context since sampling, testing, and teaching conditions can vary considerably.

Decisions based upon differences between error rates. Most of the foregoing decision rules are dependent upon the evaluator's choice of a cut-off between high and low error rate. Dichotomizing error rate in this way clearly facilitates the identification of appropriate decision rules, and, in many cases, the simplicity of the technique will probably ortweigh any loss of precision. However, we can also specify an additional pair of decision rules that take into account quantitative differences between error rates. One of these rules increases the precision of previous decisions, the other provides essentially new information. We will call these error rates "derived" error rates to distinguish them from the "raw" error rates discussed in the previous sections.

Let us consider two limitations of the high/low classification procedure for error rates. Suppose that Theoretical Error Rate (TER) and Base Error Rate (BER) for a given objective are both classified as high (H), while the Posttest Error Rate (PER) is classified as low (L). Clearly, any actual arithmetic differences between TER and BER will not affect the decisions we have thus far proposed. Also, since BER and PER are merely classified as high and low, respectively, we will not have a quantitative measure of how much learning has actually taken place.

Rules 1-3 are useful for making decisions based upon categorical differences between BER and TER, but we can make more accurate decisions by actually computing the difference between these error rates. Let

$$\text{DER} = \text{TER} - \text{BER},$$

where DER stands for "Difference Error Rate." If DER = 0, then the observed error rate on the pretest (BER) is identical to the expected error rate on the pretest (TER). If DER < 0, then fewer students are getting the item correct than we would expect on the basis of random guessing. Finally, if DER > 0, then more students are getting the item correct than we would expect on the basis of random guessing. As discussed previously, the last possibility is often an unfavorable situation, since it can mean that the item somehow "gives away" the correct answer.

We can test the significance oa a positive difference between BER and TER by computing

$$Z = \frac{\text{DER} - (1/2N)}{\sqrt{\text{TER}(1 - \text{TER})/N}},$$

where N is the total number of students in the sample. The term -1/2N is a correction for discontinuity and, as such, can be dropped if the sample size is large. Note that when TER = 1 Z is undefined; in this case, any value of DER > 0 can be considered significant. Again, however, one should be careful not assume that TER = 0 just because the format of the item is free-response. Once Z is calculated its significance can be tested by comparing the value of Z with the normal curve standard score at an appropriate α-level for a one-tailed test. Note that we are only interested in positive values of DER.

We can now specify more precise version of Rules 1-3.

Rule 8: If the value of DER is significantly greater than zero, then the item should be revised. In all other cases no revision is required.

None of the decisions discussed up to this point has made use of any measure of gain in knowledge relevant to a given objective that results from the instructional system. It is probably true that gain is not as important as final performance on the posttest, in most instructional systems; however, if students experience relatively little gain as a result of experiencing instruction, one can legitimately question the value of the instructional system itself. Thus, measures of gain have long been a subject of considerable interest in the field of instruction.

A simple measure of gain for an objective is the difference between pretest error rate (BER) and posttest error rate (PER). This measure has been suggested by Cox and Vargas (1966); however, it has one serious limitation -- gains of the same magnitude do not mean the same thing. Consider a gain of 0.50 resulting from BER = 1.00 and PER = 0.50 and a gain of the same magnitude resulting from BER = 0.50 and PER = 0.00. In the former case, the instructional system has failed to produce 50 percent of the gain in performance that could be achieved, while in the latter case, the instructional system has produced as much gain as possible given the entry level of the students. Thus, in the former case, some revision of the instruction may be desirable, while in the latter case, no revision in the instructional system is required on the basis of these data.

The above, rather trivial example, illustrates that simple gain does not provide a very meaningful basis for revising instruction. A better measure is percent of maximum possible gain for an objective defined as:

$$\text{PMPG} = \frac{\text{BER} - \text{PER}}{\text{BER}} \;.$$

In order to make use of this measure the evaluator must specify a cut-off that determines whether or not a given value of PMPG indicates a need for revision.

Rule 9: If PMPG < c, where c is a cut-off specified by the evaluator, then the instruction should be revised. The cut-off c need not bethe same for all objectives. If PMPG $\geq$ c, then, on the basis of PMPG, there is no indication that instruction needs to be revised.

The literature contains many in-depth discussions concerning the problems and pit-falls associated with measures of gain. See for example Cronbach and Furby (1970), DuBois (1962), and Harris (1963). Most of this literature, however, treats measures of gain in the context of their use in inferential statistics or correlational analysis. While we appreciate the importance of these issues, we hasten to add that measures of gain, merely as descriptive statistics, can provide useful information to evaluators. We believe that the use of PMPG, as data for evaluation purposes, is a case in point. Also, since, in criterion-referenced testing, we assume an absolute measurement scale, many of the objections to measures of gain are less crucial.

Use of Item Analysis Tables

An item analysis table indicates the number or percent of students who chose each of the alternatives of a test item. Further, in most cases, the students who responded to the item are partitioned into groups, based upon their performance on the total test. Thus, for example, if each student is put into either a "lower" or an "upper" group, then one can identify the number (or percent) of students in the lower and/or upper group who chose each alternative. Such tables, and their use in norm-referenced testing situations, are treated in practically every introductory text in educational measurement.

Item analysis tables can also be quite useful in criterion-referenced testing situations. Let us now consider some of the issues involved in constructing and using such tables.

(a) The classification of students into groups should be meaningful for the criterion-referenced test. For norm-referenced tests typical ways of partitioning students include, for example, placing the top 50 percent of the students in the upper group and the bottom 50 percent in the lower group, or partitioning students into lower, middle, and upper thirds. These procedures are not appropriate for criterion-referenced tests, because the group into which a student is classified can be determined only by referenceto the scores of other students. For criterion-referenced testing, the group into which a student is classified should be uniquely determined by the student's test score, independent of the scores of other students. In mastery testing this usually means that students who exceed the mastery cutting score are defined as the upper group of students, and all other students constitute the lower group. Thus, for criterion-referenced item analysis tables, groups are defined according to ranges of criterion-referenced or mastery test scores. In many cases, only two groups (upper and lower) are used; however, item analysis tables often provide more useful and interpretable information if one incorporates a "middle" group that contains students whose test score is on the borderline of mastery, acceptable behavior, or criterion performance.

(b) In interpreting criterion-referenced item analysis tables one should remember that if all students get all items correct, then all cells but one in every item analysis table will be empty. Furthermore, if all students get an item correct, then the only cells that will be non-empty are those associated with the correct alternative. These observations may appear trivial; however, they do emphasize an important consideration -- in criterion-referenced testing, the fact that few, or no, persons choose an incorrect alternative (distractor) does not necessarily indicate that the alternative should be revised.

(c) For the sake of discussion let D(U,L) be the difference between the proportions of students in the upper and lower groups who choose a distractor, D. One would usually expect D(U,L) to be equal to or less than zero; therefore, if D(U,L) is very much greater than zero, the di tractor or the item itself may require revision. Analogous statements referring to the correct alternative are contained in the above discussion concerning item validity and the B discrimination index.

(d) The process of analyzing criterion-referenced test items that require revision, can be conceived as a two-stage process. The first stage entails the use of decision rules such as those discussed in th previous section of this chapter; the second stage entails a detailed consideration of the item analysis table(s (It is sometimes useful to study the item analysis tables for both the pretest anu posttes': administrations of the item.) Unfortunately, at least at the present time, the use of item analysis tables is probably more an art than it is a science. Nevertheless, careful subjective analysis of item analysis tables will often reveal the presence of problems that are not apt to be evident from the typical kinds of descriptive statistics for items.

CHAPTER VI

An Alternative to the Classical Administration and Scoring Procedure For Analyzing Criterion-Referenced Test Items

In Chpater V we considered, in some detail, procedures for analyzing criterion-referenced test items when students are forced to pick one and only one alternative and scored either correct (1) or incorrect (0). With very few exceptions, researchers in the field of criterion-referenced testing have concerned themselves only with this classical procedure for the administration and scoring of items

For norm-referenced testing classical correct/incorrect administration and scoring procedures seem to be reasonably effective and useful. However, norm-referenced tests are usually relatively long; the scores from such tests are often normally distributed; floor and ceiling effects seldom occur in norm-referenced tests; and, most importantly, one is not very much concerned about the precise proportion of items a student can answer correctly -- rather, one is concerned about the ability of the test to distinguish among subjects. Each of these characteristics of a norm-referenced test argues directly or indirectly that the classical correct/incorrect procedure is reasonably adequate (or, at least, not grossly inadequate) for many norm-referenced tests.

On the other hand, criterion-referenced tests are usually short; the scores from such tests are often negatively skewed -- even severely so; ceiling effects are very common; and, most importantly, one is fundamentally concerned about accurately estimating the proportion of items to which a student knows the answer (or possibly some other score). This emphasis on accurate estimation of a student's score is especially critical in criterion-referenced testing because there is seldom any external criterion measure for judging validity.

Thus, in criterion-referenced testing it is very important to use every possible means of eliminating random (and systematic) errors of measurement . In particular, it seems to this author that it is important to eliminate (or, at least, be able to estimat the effect of) guessing. Now, it is very clear that, a considerable amount of student guessing frequently occurs when a student if forced to pick one and only one alternative and the classical correct/incorrect scoring procedure is used; moreover, when the classical procedure is used, it is very difficult, if not impossible, to ascertain the magnitude of the effect of guessing upon student scores.

Furthermore, since criterion-referenced tests are frequently short, it seems desirable to obtain as much informations as possible from each item; yet, using the classical procedure for administering and scoring an item, one merely knows whether or not the student got the item correct. In particular, using the classical procedure one does not obtain information with regard to the relative attractiveness of each alternative for each student. This kind of information can be very useful in determining whether or not to revise a criterion-referenced test item. Thus, the classical procedure somewhat limits the amount of information we obtain with regard to any given criterion-referenced test item.

In short, from a criterion-referenced testing viewpoint, this author feels that the classical procedure for administering and scoring an item has two serious limitations: (a) scores obtained using this procedure incorporate an indeterminable amount of guessing and (b) this procedure provides very little information with regard to any given item especially when relatively small numbers of students take the item. These points imply that when we use the classical procedure for criterion-referenced testing, we may have less than adequate information for determining whether or not a criterion-referenced test item requires revision.

Therefore, it is worthwhile to consider alternatives to the classical procedure. There are a number of points of view from which one could consider different procedures Here we are interested in the ability of the procedure to aid us in item analysis. That is, our goal is to identify a procedure for administering an item that provides us with optimum data for determining whether or not the item needs to be revised; and, if possible, these data should aid us in pinpointing the nature of any difficulties with the item. For this purpose, we consider two potential procedures which we call the "elimination procedure" and the "confidence procedure." We find that the confidence procedure is the better of the two for our purposes.

It should be noted that here we are not concerned about the kinds of scores typically obtained from the elimination and confidence procedures; rather, our primary concern is with the nature and amount of data collected when such procedures are used. Also, we do not assume that once an item is administere. using one procedure it will always be administered using that procedure. .In fact, when we consider the confidence procedure, the manner in which we interpret the data provides us with a kind of guessing-free estimate of a person's classical score. Thus, once an item has been validated using the confidence procedure, one can administer the item using the classical procedure.

Two Alternatives to the Classical Procedure for Administering Items

Elimination procedure. Coombs et al (1956) suggest a procedure for administering and scoring a test based upon having students eliminate alternatives that they consider to be incorrect. Since a student may eliminate any number of alternatives for any test item, the elimination procedure provides some information about the relative attractiveness of each alternative. However, the information provided is somewhat ambiguous in that, for example, if a student eliminates two alternatives, we so not know whether or not the student feels more uncertain about one alternative than the other.

Also, let us consider the elimination procedure from another point of view. As indicated previously, we are interested in a procedure's ability to provide us with a kind of guessing-free estimate of a person's classical score. Let us call such an estimate a PC1 score, indicating the probability (P) that a person's classical (C) score on an item is unity (1). If we know, for example, that a person guessed randomly on a four-alternative item, then PC1 should be 0.25. The question is, "Can the kind of data collected using the elimination procedure provide us with an adequate basis for estimating a student's PC1 score for an item?"

Suppose, for example, that a student eliminates two alternatives for a four-alternative item. If we could assume that, when forced to pick one and only one alternative, the student would randomly pick one of the two non-eliminated alternatives, then the PC1 score for the student for the item would be 0.50. However, this assumption is not necessarily valid; in fact, one could argue that PC1 might be any value between 0.50 and 1.00.

Thus, it does not appear that the elimination procedure provides an adequate basis for estimating a student's PC1 score for an item. Consequently, if the student were administered the item a large number of times, we don't have a very good basis for estimating the number, or proportion, of times the student would get the item correct under the classical scoring procedure. If the item is administered K times, this proportion should be K·PC1.

Confidence procedure. In confidence testing, one obtains from each student a subjective probability that each alternative of a test item is correct. There are a number of techniques that can be used to obtain these probabilities either directly or indirectly. This author prefers the technique usually called the "star" method in which a student is told to distribute a fixed number of "stars" or points over the alternatives of a test item. For example, students might be told to distribute twelve points over the alternatives of a four-alternative item. The table below indicates some of the ways students might perform this task and the associated (subjective) probabilities.

	No. of Points				Probabilities				
	A*	B	C	D	A*	B	C	D	PC1
S_1	3	3	3	3	.25	.25	.25	.25	.25
S_2	4	4	[illegible]	0	.33	.33	.33	.00	.33
S_3	6	6	0	0	.50	.50	.00	.00	.50
S_4	12	0	0	0	1.00	.00	.00	.00	1.00
S_5	5	5	1	1	.42	.42	.08	.08	.50
S_6	5	2	4	1	.42	.17	.33	.08	1.00

The reader interested in a more in-depth discussion of confidence testing can consult de Finetti (1965), Echternacht (1972), Savage (1971), and Shuford et al (1966).[1] A great deal of the literature on confidence testing involves discussion of various procedures for scoring such tests, but this is not our concern in this chapter.

[1] Appendix A to this report is a manual for DEC-TEST, a computer program that analyzes confidence test data in great detail. Further, the introduction to this manual provides a description of confidence testing and elimination testing as these procedures are typically used.

Here we are concerned about the nature of the data (i.e., the probabilities) collected for each item and for each student.

Each probability indicates how confident the student is that the particular alternative is the correct answer for the item. Using these probabilities we can obtain PC1 scores from the following rules:

Let M = the magnitude of the highest probability for a particular student for a given item,

A = the number of alternatives for the item,

$P(a)$ = the probability associated with alternative a ($a = 1, 2, \ldots, A$), and

$*$ = the correct alternative.

Now,

PC1 = 0 if $P(*) \neq M$;

PC1 = $1/K$ if $P(*) = M$ and there are $(K-1)$other alternatives having $P(a) = M$; and

PC1 = 1 if $P(*) = M$ and there are not other alternatives having $P(a) = M$.

See the table on the previous page for examples of PC1 scores. Note, in particular, that the third and fifth students both have PC1 = 0.50 eventhough $M = 0.50$ for the third student and $M = 0.42$ for the fifth student.

Thus, PC1 scores are readily available from the subjective probabilities one obtains using the confidence testing procedure. Furthermore, when one uses confidence testing as a procedure to collect data for items, one obtains, for each student, a probability associated with each alternative for each item. Thus, one has a great deal of information for each item -- much more information than if students pick one alternative or eliminate alternatives.

In short, the confidence procedure seems to be superior to the elimination procedure, at least for out purposes here.

Item Analysis Tables from the Confidence Procedure

Conisder the synthetic data for a hypothetical item presented in Table 6-1. The item has four alternatives, "A" is the correct answer, and the twenty students are partitioned into lower and upper groups of ten students each. The confidence probabilities are indicated for each alternative and for each student. We emphasize that these are synthetic data, and they are not necessarily indicative of a good criterion-referenced test item, we use these data merely to illustrate our discussion.

For each confidence probability in Table 6-1, there is a pseudo-classical score. A pseudo classical score for an alternative is defined as the probability that a student would pick the alternative if the student were forced to choose one and only one alternative for the item under consideration. Thus, the pseudo-classical score for an item is the pseudo-classical score for the correct alternative; also, the pseudo-classical score for an item is identical to the PC1 score discussed previously.

Using the data in Table 6-1, one can construct the item analysis tables given by Tables 6-2 and 6-3, where Table 6-2 uses confidence probabilities and Table 6-3 uses pseudo-classical scores. Both tables present frequency distributions of scores on alternatives, with associated totals, means, and standard deviations. Clearly, Table 6-2 provides more information, and a somewhat different kind of information than Table 6-3; and, both tables provide much more information than is available from item analysis tables based upon the classical correct/incorrect scoring procedure. This additional information can be quite useful in deciding what (if anything) is wrong with a criterion-referenced test item.

Now, let us summarize a few points implicit in our discussion thus far. We are assuming that once an item is validated it probably will be administered using the classical correct/incorrect scoring procedure. However, in order to validate the item we are suggesting that the evaluator collect confidence probabilities for each alternative, translate these probabilities to pseudo-classical scores for each alternative, and generate the pseudo-classical item analysis table. This table indicates the probability the each student would pick each alternative using the classical correct/incorrect scoring procedure; thus, using this table one can analyze the probable effect of guessing upon the performance of other similar students who take the item using the classical procedure for item administration and scoring. Further,

TABLE 6-1

Synthetic Data

	Student No.	Confidence Probabilities				Pseudo-classical scores[a]			
		A*	B	C	D	A*	B	C	D
Lower Group	1	.25	.25	.25	.25	.25	.25	.25	.25
	2	.25	.25	.25	.25	.25	.25	.25	.25
	3	.40	.40	.10	.10	.50	.50	.10	.10
	4	1.00	.00	.00	.00	1.00	.00	.00	.00
	5	.30	.20	.30	.20	.50	.00	.50	.00
	6	.50	.50	.00	.00	.50	.50	.00	.00
	7	.30	.30	.10	.30	.33	.33	.00	.33
	8	.20	.70	.00	.10	.00	1.00	.00	.00
	9	.40	.20	.00	.40	.50	.00	.00	.50
	10	.00	1.00	.00	.00	.00	1.00	.00	.00
	Sum-L[b]	3.60	3.80	1.00	1.60	3.83	3.83	1.00	1.33
	Mean-L	.36	.38	.10	.16	.38	.38	.10	.13
	SD-L	.25	.27	.12	.13	.28	.45	.17	.18
Upper Group	11	.25	.25	.25	.25	.25	.25	.25	.25
	12	1.00	.00	.00	.00	1.00	.00	.00	.00
	13	1.00	.00	.00	.00	1.00	.00	.00	.00
	14	.70	.20	.00	.10	1.00	.00	.00	.00
	15	.60	.00	.20	.20	1.00	.00	.00	.00
	16	.50	.50	.00	.00	.50	.50	.00	.00
	17	.40	.50	.00	.10	.00	1.00	.00	.00
	18	.50	.50	.00	.00	.50	.50	.00	.00
	19	.80	.10	.10	.00	1.00	.00	.00	.00
	20	.30	.30	.30	.10	.33	.33	.33	.00
	Sum-U[b]	6.05	2.35	.85	.75	6.58	2.58	.58	.25
	Mean-U	.61	.24	.09	.08	.66	.26	.06	.03
	SD-U	.25	.20	.11	.09	.37	.32	.12	.08
	Sum-T[b]	9.65	6.15	1.85	2.35	10.41	6.41	1.58	1.58
	Mean-T	.48	.31	.09	.12	.52	.32	.08	.08
	SD-T	.28	.25	.12	.12	.12	.34	.15	.15

[a] A pseudo-classical score for an alternative represents the probability that a student would pick the alternative if the student were forced to pick one and only one alternative for the test item.

[b] L, U, and T mean the lower, upper, and total groups, respectively.

TABLE 6-2

Item Analysis Table Using Confidence Probabilities

Probability Interval	A*			B			C			D		
	Low	Up.	Tot	Low	Up.	Tot	Low	Up.	Tot	Low	Up.	Tot
$0.0 \le P < 0.1$	1	0	1	1	3	4	5	6	11	3	5	8
$0.1 \le P < 0.2$	0	0	0	0	1	1	2	1	3	2	3	5
$0.2 \le P < 0.3$	3	1	4	4	2	6	2	2	4	3	2	5
$0.3 \le P < 0.4$	2	1	3	1	1	2	1	1	2	1	0	1
$0.4 \le P < 0.5$	2	1	3	1	0	1	0	0	0	1	0	1
$0.5 \le P < 0.6$	1	2	3	1	3	4	0	0	0	0	0	0
$0.6 \le P < 0.7$	0	1	1	0	0	0	0	0	0	0	0	0
$0.7 \le P < 0.8$	0	1	1	1	0	1	0	0	0	0	0	0
$0.8 \le P < 0.9$	0	1	1	0	0	0	0	0	0	0	0	0
$0.9 \le P \le 1.0$	1	2	3	1	0	1	0	0	0	0	0	0
Total[a]	3.60	6.05	9.65	3.80	2.35	6.15	1.00	.85	1.85	1.60	.75	2.35
Mean[a]	.36	.61	.48	.38	.24	.31	.10	.09	.09	.16	.08	.12
Stan Dev.[a]	.25	.28	.28	.27	.20	.25	.12	.11	.12	.13	.09	.12

[a] These statistics are based upon the actual value of each confidence probability; they are not based upon the midpoint of the probability interval within which the confidence probability lies. See Table 6-1.

TABLE 6-3

Item Analys Table Using Pseudo-classical Scores

Pseudo-classical Score	A*			B			C			D		
	Low	Up.	Tot	Low	Up.	Tot	Low	Up.	Tot	Low	Up.	Tot
0.00	2	1	3	3	5	8	7	8	15	6	9	15
0.25	2	1	3	2	1	3	2	1	3	2	1	3
0.33	1	1	2	1	1	2	0	1	1	1	0	1
0.50	4	2	6	2	2	4	1	0	1	1	0	1
1.00	1	5	6	2	1	3	0	0	0	0	0	0
Total[a]	3.83	6.58	10.41	3.83	2.58	6.41	1.00	.58	1.58	1.33	.25	1.58
Mean[a]	.38	.66	.52	.38	.26	.32	.10	.06	.08	.13	.03	.08
Stan Dev[a]	.28	.37	.12	.45	.32	.34	.17	.12	.15	.18	.08	.15

[a] These statistics are based upon the actual value of each pseudo-classical probability.

if one wants a detailed display of the certainty with which students choose any alternative, one can generate the item analysis table based upon the confidence probabilities.

Admittedly, the ideas discussed above require detailed procedures for item administration, scoring, and analysis; however, the additional time and effort required can, I think, be very worthwhile for the process of validating items.

An Application of PC1 Scores in the Classical Test Theory Model

Recall that under the classical test theory model $X = T + E$, where X, T, and E are observed, true, and random error scores, respectively. Now, we have described the PC1 item score for a student as a kind of guessing-free estimate of a person's classical score, and guessing is usually interpreted as one kind of random error. If we assume that guessing is the only, or the principal, kind of random error that concerns us, then a PC1 score is a kind of true score and we can analyze the effect of guessing upon classical scores by using the classical test theory model directly. Thus, in this section we will let

$X = 0$ or 1 (classical observed score),

$T =$ PC1 item score, and

$E =$ random error due to guessing.

Basic statistics. Note that when one typically uses the classical test theory model, one has observed scores, and one wants to estimate true scores; however, in this case, we already have the true scores, and we must estimate the observed scores. Now, if the item were administered to student i a total of K times we would expect student i to get the item correct $K \cdot T_i$ times, and we would expect student i to get the item incorrect $K \cdot (1-T_i)$ times. Therefore, if N is the total number of subjects

$$\bar{X} = \frac{1}{KN} \sum_{i=1}^{N} K \cdot T_i \qquad (6.1)$$

$$= \bar{T}$$

$$\text{and } s_X^2 = \frac{1}{KN}\sum_{i=1}^{N} K\cdot T_i - \bar{T}^2$$

$$= \bar{T} - \bar{T}^2$$

$$= \bar{T}(1 - \bar{T}) \quad . \qquad (6.2)$$

For an example of these statistics see Table 6-4 which uses the synthetic data presented in Table 6-1 and assumes, for the sake of illustration, that X = 12.

Table 6-4 also indicates the error scores associated with each observed score for our synthetic data. The mean and variance of the error scores are given by:

$$\bar{E} = \frac{1}{KN}\sum_{j=1}^{K}\sum_{i=1}^{N}(X_{ij} - T_{ij})$$

$$= \frac{1}{KN}\sum_{i=1}^{N} K\cdot T_i - \frac{1}{N}\sum_{i=1}^{N} T_i$$

$$= \bar{T} - \bar{T}$$

$$= 0 \qquad (6.3)$$

$$\text{and } s_E^2 = \frac{1}{KN}\sum_{j=1}^{K}\sum_{i=1}^{N}(X_{ij} - T_{ij})^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{K}\sum_{j=1}^{K}(X_{ij} - T_{ij})^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{K}\sum_{j=1}^{K} X_{ij} - \frac{2}{K}\sum_{j=1}^{K} X_{ij}T_{ij} + \frac{1}{K}\sum_{j=1}^{K} T_{ij}^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[T_i - \frac{2}{K}(K\cdot T_i^2) + \frac{1}{K}(K\cdot T_i^2)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}(T_i - T_i^2)$$

$$= \frac{1}{N}\sum_{i=1}^{N} T_i(1 - T_i) \qquad (6.4)$$

TABLE 6-4

Observed, True, and Error Scores

Student	Dist. of Observed Scores		True Scores	Frequency Distribution of Error Scores						
	1	0		-0.50	-0.33	-0.25	0.00	0.50	0.67	0.75
1	3	9	0.25			9				3
2	3	9	0.25			9				3
3	6	6	0.50	6				6		
4	12	0	1.00				12			
5	6	6	0.50	6				6		
6	6	6	0.50	6				6		
7	4	8	0.33		8				4	
8	0	12	0.00				12			
9	6	6	0.50	6				6		
10	0	12	0.00				12			
11	3	9	0.25			9				3
12	12	0	1.00				12			
13	12	0	1.00				12			
14	12	0	1.00				12			
15	12	0	1.00				12			
16	6	6	0.50	6				6		
17	0	12	0.00				12			
18	6	6	0.50	6				6		
19	12	0	1.00				12			
20	4	8	0.33		8				4	

$\overline{X} = 0.521$ $\quad \overline{T} = 0.521$ $\quad \overline{E} = 0.000$

$s_X^2 = 0.249$ $\quad s_T^2 = 0.124$ $\quad s_E^2 = 0.125$

Now, let us demonstrate that $s_X^2 = s_T^2 + s_E^2$.

$$s_T^2 + s_E^2 = [\frac{1}{N}\sum_{i=1}^{N} T_i^2 - \bar{T}^2] + [\frac{1}{N}\sum_{i=1}^{N} T_i(1 - T_i)]$$

$$= \frac{1}{N}\sum_i T_i^2 - \bar{T}^2 + \frac{1}{N}\sum_i T_i - \frac{1}{N}\sum_i T_i^2$$

$$= \bar{T} - \bar{T}^2$$

$$= \bar{T}(1 - \bar{T})$$

$$= s_X^2 .$$

Thus, we have demonstrated that, by interpreting our PC1 scores as true scores we can express the mean and variance of observed scores in terms of the true scores. Furthermore, we have shown that the variance of the observed scores does indeed equal the variance of the true scores plus the variance of the error scores. The mean and variance of the observed, true, and error scores are provided in Table 6-4. For reference now and later, the reader should note that, for our synthetic data

$$\sum_{i=1}^{20} T_i = 10.41 ,$$

$$\sum_{i=1}^{20} T_i^2 = 7.9053 , \text{ and}$$

$$\sum_{i=1}^{20} T_i^3 = 6.8687 .$$

Reliability of a one-item test. Using the above results, we can express the reliability of a one-item test as:

$$r_{11} = s_T^2 \; / \; s_X^2$$

$$= \frac{\dfrac{\Sigma T^2}{N} - \bar{T}^2}{\bar{T}(1 - \bar{T})}$$

$$= \frac{\Sigma T^2 - N \cdot \bar{T}^2}{\Sigma T - N \cdot \bar{T}^2} \qquad (6.5)$$

For our synthetic data,

$$r_{11} = \frac{0.124}{0.249} = 0.498 \quad .$$

The reader should keep in mind that r_{11} is the proportion of variance in observed scores not due to guessing, whereas $(1 - r_{11})$ is the proportion of variance in observed scores due to guessing. Now, we call r_{11} the reliability of a one-item test; however, if there are random errors operating other than those due to guessing, then r_{11} will be an upper- limit to the "true" reliability of the item.

In order to estimate the reliability of a test consisting of K replications of the item, we can use the Spearman-Brown Prophecy Formula

$$r_{KK} = \frac{Kr_{11}}{1 - (K - 1)r_{11}} \qquad (6.6)$$

Another way to view the reliability of a one-item test is to ask how many items of a similar nature would have to be administered in order to obtain a given level of reliability. This question can be answered by re-arranging the terms in the Spearman-Brown Prophecy Formula in order to get

$$K = \frac{r_{KK}(1 - r_{11})}{r_{11}(1 - r_{KK})} \quad ., \qquad (6.7)$$

where, in this case, r_{KK} is the level of reliability desired and K is the number of items necessary to achieve this level of reliability. Using our synthetic data, if we set r_{KK} = 0.90, then

$$K = \frac{0.90(1 - 0.498)}{0.498(1 - 0.90)} = 9.072 \ .$$

One further statistic, of a reliability nature, may be of interest. It can be shown that the probability that a randomly selected student would maintain his or her observed score on L = 2 or 3 administrations of the same item is:

$$P_L = 1 - s_E^2 \ . \qquad (5.8)$$

For our synthetic data,

$$P_2 = 1 - 2(0.125) = 0.750$$

and $P_3 = 1 - 3(0.125) = 0.635$.

Regression of observed scores on true scores. The standard error of measurement is the square root of the expression in (6.4), which is also equal to

$$s_E = s_X\sqrt{1 - r_{11}} \qquad (6.9)$$

For our synthetic data,

$$s_E = \sqrt{0.125} = 0.354$$

or $s_E = \sqrt{0.249}\,\sqrt{1 - 0.498} = 0.354$.

The reader should recall that the standard error of measurement is associated with the regression of observed scores on true scores, as indicated, for our synthetic data, in Figure 6-1. This regression is used to predict observed scores from true scores. As such, this regression can be used to establish a confidence interval around the expected difficulty level of the item, where difficulty level is based on the classical scoring procedure and is merely the proportion of subjects who get an item correct.

Regression of true scores on observed scores. The other regression of interest is the regression of true scores on observed scores. From classical test theory, this regression is:

$$\hat{T} = \bar{T}(1 - r_{11}) + r_{11}X \tag{6.10}$$

where $\hat{T}$ is the estimated value of T assuming a linear regression of true on observed scores. The standard deviation of errors about this regression is called the standard error of estimate and denoted s_{est}. For the kind of data considered here, it can be shown that

$$\begin{aligned} s^2_{est} &= \left[\frac{\Sigma T - \Sigma T^2}{N} \right] \left[\frac{N\Sigma T^2 - (\Sigma T)^2}{N\Sigma T - (\Sigma T)^2} \right] \qquad (6.11) \\ &= s^2_E\, r_{11} \end{aligned}$$

Now, since there are only two possible observed scores for an item (0 and 1) it is also true that

$$s^2_{est} = w_0 s^2_{est(0)} + w_1 s^2_{est(1)} \text{ , where} \tag{6.12}$$

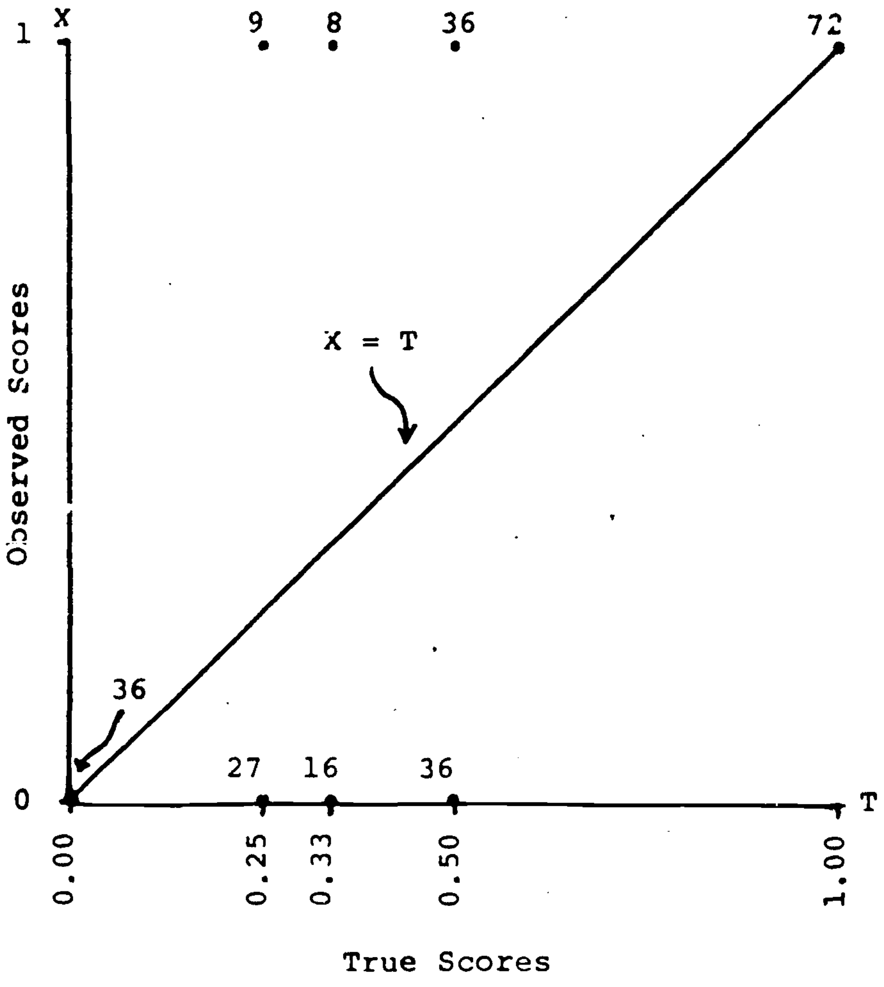
$s^2_{est(0)}$ = the variance of the errors about the regression line when X = 0

$$= \left[\frac{\Sigma T^2 - \Sigma T^3}{N - \Sigma T} \right] - \left[\frac{\Sigma T - \Sigma T^2}{N - \Sigma T} \right]^2 , \tag{6.13}$$

$$w_0 = 1 - \bar{T} , \tag{6.14}$$

FIGURE 6-1

Regression of Observed Scores on True Scores



$s_E^2 = 0.125$

$s^2_{est(1)}$ = the variance of the error scores about the regression line when X = 1

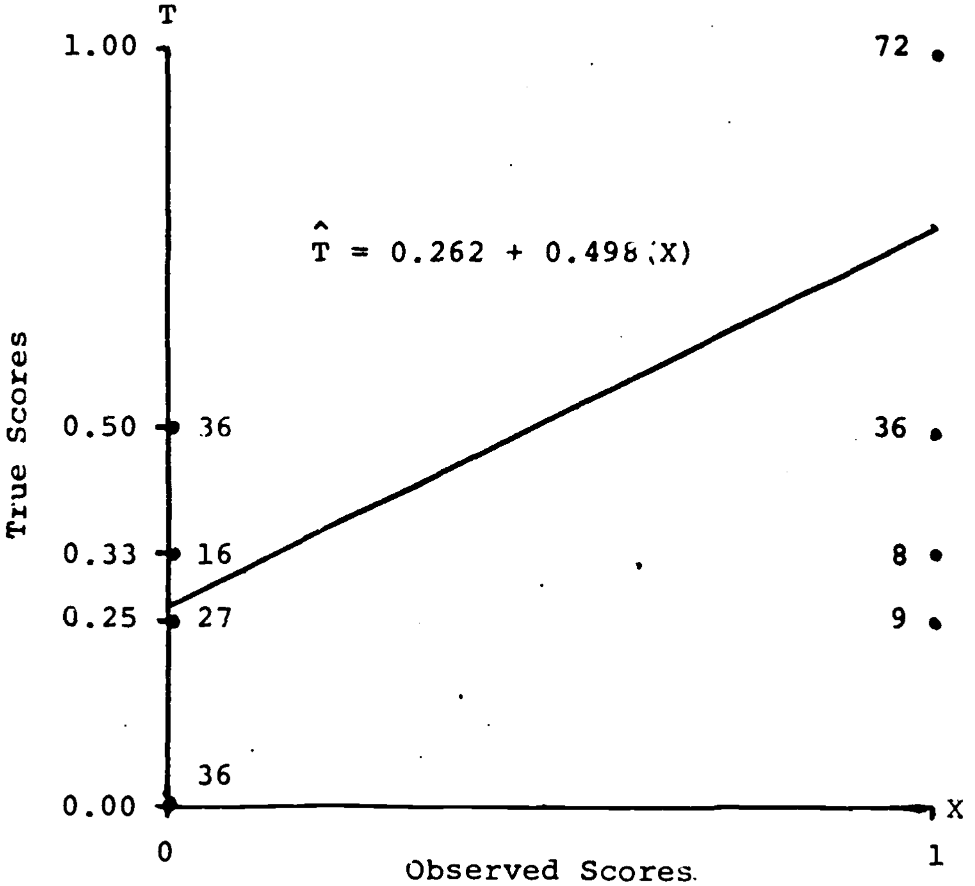
$$= \frac{\Sigma T^3}{\Sigma T} - \left[\frac{\Sigma T^2}{\Sigma T}\right]^2 \text{, and} \qquad (6.15)$$

$$w_1 = \bar{T} \ . \qquad (6.16)$$

Figure 6-2 provides, for our synthetic data, the regressions of true scores on observed scores, as well as the values of the statistics indicated in (6.11), (6.13), and (6.15).

FIGURE 6-2

Regression of True Scores on Observed Scores



$s^2_{est} = 0.062$

$s^2_{est(0)} = 0.040$

$s^2_{est(1)} = 0.083$

CHAPTER VII

Data Analysis

In this chpater we define, present, and discuss a set of data that were collected in order to illustrate some of the issues, statistics, and procedures considered in previous chapters. The data reported should not be considered as necessarily indicative of either "good" or "bad" criterion-referenced tests or items.

Design for Data Collection

In the fall of 1972 and the spring of 1973 two forms (A and B) of a 25-item criterion-referenced test for a course in educational measurement were administered in both the pre- and posttest mode to 113 students.

In order to understand the design used for administering these tests, the reader will find it useful to refer to the format of Table 7-la[1]. In this table (and other tables to be discussed in this chapter) the following notation is used:

Factor	Level	Description
A[2]	a_1	test administered using SCoRule
A[2]	a_2	test administered using "star" technique
B[2]	b_1,b_2,b_3,b_4	blocks of subjects
C	c_1	Form A of test
C	c_2	Form B of test
D	d_1	Pretest
D	d_2	Posttest

Also, note that a "." in place of a subscript indicates the mean over all levels of the factor being considered.

[1] All tables referenced in this chapter can be found at the end of the chapter.

[2] Factors A and B should not be confused with forms A and B of the Pretest and the Posttest.

The reader should note several important facts about this design:

(a) If we collapse the levels of the A factor, we see that subjects in the first block received Pretest A and Posttest A, subjects in the second block received Pretest A and Posttest B, subjects in the third block recieved Pretest B and Posttest A, and subjects in the fourth block received Pretest B and Posttest B. Furthermore, note that subjects were randomly assigned to blocks.

(b) The discussion above indicates that the design is a (balanced) repeated measures design in which half of the available cells are empty· i.e., each subject took one form of the Pretest and one form of the Posttest, and, thus, no subject took both forms of either the Pretest or the Posttest. In the opinion of this author, the constraints incorporated in the design are realistic in that it is often not feasible to obtain repeated measures for equivalent tests in the real world of course development and evaluation.

(c) Although the constraint mentioned above is realistic, it is, nevertheless, somewhat restricting. For example, we cannot obtain direct measures of the equivalence of the two forms of the Pre- and Posttests. Also, when we examine summary statistics for tests and items, these statistics sometimes will be based upon different or partially overlapping samples of subjects.

The actual items administered to subjects are provided in Appendix B (see footnote 1, below). All items are four-alternative objective items which had not been subjected to any previous validation or revision procedures. Therefore, these items are not necessarily "good" items. In fact, one of the purposes of this chapter is to illustrate a procedure discussed in Chapter V that might be used to collect data, report statistics, and identify items that may require revision. All test and item data were analyzed using DEC-TEST, which is described in Appendix A, and SPSS.

[1] Note that Forms A and B of the Posttest actually contained 50 items; however, items 26-50 (identified as ZC26 to ZC50 in Appendix B) were the same items in both forms, and none of these items was intended to be equivalent to any item numbered 1 to 25. Therefore, for the purposes of this chapter, we shall treat only items 1 to 25.

Another important aspect of the data collection procedure involves the way in which students responded to test items. For each item, each student identified the alternative he or she would pick if forced to pick one and only one alternative; also, each student indirectly reported his or her subjective probabilities for each alternative for each item. Subjects in level a_1 reported actual log scores (range of 0 to 100) for each alternative using a mechanical device called a SCoRule; these log scores were later transformed into subjective probabilities using a formula provided in Appendix A (see p. A-28). Students in level a_2 used the twelve-point "star" system for reporting their subjective probabilities (see Chapter VI and/or p. A-27). The reader unfamiliar with confidence testing, the logarithmic scoring system, subjective probabilities, and/or the "star" system would be well-advised to study pages 6-1 to 6-10, and the first section of Appendix A.

Summary Statistics for Subjects and Tests

The procedure whereby subjects responded to items may be summarized by saying that subjects did two things -- they picked one alternative and they indirectly reported subjective probabilities. The "pick one" procedure allows us to calculate a classical correct/wrong (1 or 0) item score for each subject, while the "subjective probability" procedure (typically considered in conjunction with confidence testing, admissible probability measurement, or decision-theoretic testing) allows us to calculate or estimate a number of different item scores for each subject. (See Section I of Appendix A, especially pages A-11 to A-14.)

Tables 7-1a,b,c to 7-6a,b,c report means and standard deviations over tests and persons for six different types of subject scores. In these and other tables, the different scores for a subject are identified as:

VAR(1) = Arithmetic mean of item confidence scores; i.e., each subject's score is the arithmetic mean of the subjective probabilities associated with the correct answer to each item. (Range = 0 to 1.)

VAR(2) = Geometric mean of item confidence scores; i.e., each subject's score is the geometric mean of the subjective probabilities associated with the correct answer to each item. See Appendix A, p. A-3 for formulas. (Range = 0 to 1.)

VAR(3) = Arithmetic mean of item log scores; i.e., each subject's score is the arithmetic mean of the log scores associated with the correct answer to each item. (Range = 0 to 100.)

VAR(4) = Arithmetic mean of item elimination scores, which are estimated from the subject's subjective probabilities using a procedure described in Chapter VI, p. 6-3, and Appendix A, pp. A-12 to A-14. (Range = -1 to 1.)

VAR(5) = Arithmetic mean of item pseudo-classical scores, which are estimated from the subject's subjective probabilities using the procedure described in Chapter VI, pp. 6-4 to 6-5, and Appendix A, p. A-14. (Range = 0 to 1)

VAR(6) = Arithmetic mean of classical item scores, which are determined directly from the "pick one" procedure. (Range = 0 to 1.)

Table 7-7 reports means, standard deviations, and reliabilities for each of the four tests and for each of the six different kinds of subject scores. The reader should note that we report these reliabilities mainly for the sake of completeness. We do not claim that any of these tests consist of a homogeneous set of items, which is a logical pre-requisite to a meaningful internal consistency reliability.

Tables 7-1 to 7-7 are presented for the reader who is interested in comparing the six different types of scores discussed above. For our purposes, in this chapter, we will concentrate primarily upon pseudo-classical scores. Recall that pseudo-classical scores are estimated classical scores which are determined from the subjective probabilities assigned by subjects to the alternatives of test items. As indicated previously, pseudo-classical scores are much less affected by guessing than are classical scores, one can directly determine a kind of item reliability from pseudo-classical item scores, and pseudo-classical scores are easily interpreted. Pseudo-classical scores , in fact, appear to have most of the advantages and few of the disadvantages of both classical scores and subjective probabilities.

In short, in the opinion of this author, pseudo-classical scores have considerable promise as a basis for validating criterion-referenced, mastery, and possibly norm-referenced test items. It should be noted that once an item has been validated using pseudo-classical scores, one can logically consider subsequently administering and scoring the validated item using classical procedures; however, it is somewhat more difficult to justify validating an item using log scores, subjective probabilities, or elimination scores, and then subsequently administering and scoring the item using classical procedures.

Test means and standard deviations using pseudo-classical scores are presented in Tables 7-5a,b,c. Note that Tables 7-5b and 7-5c are primarily different ways of displaying the data in the cells of Table 7-5a. Let us now consider three hypotheses for both the Pretest (Table 7-5b) and the Posttest (Table 7-5c) aspects of these data:

(a) There are no differences among means for those subjects who used the SCoRule (level a_1) versus those subjects who used the star technique (level a_2) for recording responses to items.

(b) The two forms of the test have equal means.

(c) There are no differences among the means for subjects in each of the four blocks. Recall that subjects were randomly assigned to blocks, and, therefore, we would not expect to find any such differenc

The results of testing these hypotheses are indicated in Tables 7-8 and 7-9, which are based upon the data in Tables 7-5b and 7-5c, respectively. (See footnote 1, below.) In both Tables 7-8 and 7-9, the first six contrasts are related to the first hypothesis, the seventh contrast is related to the second hypothesis, and the last two contrasts are related to the third hypothesis. All contrasts were defined a priori.

[1] In Tables 7-8 and 7-9, the columns labelled "orth t" and "Bonf t" provide an indication of significance levels for multiple comparisons using the orthogonal t-test procedure and the Bonferroni t-test procedure, respectively. (The latter is also called Dunn's procedure.) Strictly speaking, for these analyses, the orthogonal t-test procedure is too liberal in declaring significant differences. The Bonferroni procedure is more conservative.

Let us now examine what the data reveal about each hypothesis:

(a) There is a significant main effect for Factor A on the Pretest but not on the Posttest. One half of the Pretest contrasts that compare levels of A are significantly different from zero using the orthogonal t-test procedure for multiple comparisons. For all but one of the Pretest and Posttest contrasts, there is an indication that students in level a_2 achieved higher scores than students in level a_1. In short, there is a definite trend for subjects who use the star technique to achieve higher scores than those who use the SCoRule, and this trend is more pronounced on the Pretest than on the Posttest. Probably these results indicate that students understand the star technique better than they understand the use of the SCoRule.

(b) Contrast number seven in Tables 7-8 and 7-9 indicates that the difference between the means for Forms A and B, for both the Pretest and the Posttest, is not significant. The reader should note, however, that differences between forms are confounded with differences between blocks. The best we can say is that we have no direct evidence to reject the hypothesis that forms are equivalent.

(c) There is a significant main effect for Factor B on the Posttest but not on the Pretest. Contrast number nine in Table 7-9 indicates that the significant Posttest difference is primarily a result of the difference between the means for subjects in the second and fourth blocks (i.e., subjects who took Posttest B). Since subjects were randomly assigned to blocks, the author has no explanation for this result, other than the rather obvious statement that random assignment does not guarantee equality of means. (Note that Table 7-5c indicates that block b_2 for the Posttest has a considerably lower mean than any other block for the Posttest, including blocks associated with Posttest A.)

In the next two sections we will analyze each of the items that make up both forms of the criterion-referenced Pretest and Posttest. In these sections we will continue to emphasize pseudo-classical item scores, although we will, on occasion, report statistics based upon subjective probabilities associated with items and classical item scores.

Item Equivalence

Let us review the nature of each of the tests considered here. There are two forms (A and B) of the Pretest and two forms (A and B) of the Posttest. Pretest A and Posttest A are identical, item by item, and the same is true of Pretest B and Posttest B. If we let "i" be a generic item number, then item i on Form A (in both the Pre- and Posttest) is intended to be equivalent to item i on Form B (in both the Pre- and Posttest). In brief, there are two different tests, or sets of items (Form A and Form B) administered at two different times (Pretest and Posttest). Consequently, a complete analysis of item equivalence must consider the issue of equivalence for each item for both the Pretest and Posttest mode.

If we generalize from classical procedures for testing the equivalence of two tests, we would test the equivalence of two items in, say, the Posttest mode, by administering both items to the same set of subjects at the time of the Posttest. Then, if the means and standard deviations of the two items were the same, we could claim that the two items are equivalent, and the correlation between the item scores for the two items could be interpreted as a coefficient of equivalence for the item. However, the design used to collect our data will not permit such a procedure since, as indicated previously, the same subjects never take both forms of an item in either the Pretest or the Posttest mode. Also, for this reason, we cannot use Cochran's Q-test (discussed in Chapter V, p. 5-7) to test item equivalence when items are scored in the classical correct/wrong manner.

In short, we cannot obtain a direct measure of item equivalence for the two forms of any item given the design for data collection employed here. However, since subjects were randomly assigned to blocks, and since, for the most part, there are no significant differences between block means for the Pre- and Posttests, we can partially consider the statistical issue of item equivalence by examining the differences between Form A and Form B item means and standard deviations. Tables 7-10 to 7-12 present the appropriate item statistics when items are scored using subjective (confidence) probabilities, classical scores, and pseudo-classical scores, respectively.

Let us consider Table 7-12, which is based upon speudo-classical item scores, in some detail. The means reported can be interpreted in a manner similar to item difficulty levels. The difference between means for the two forms of any item is tested using a t-test for independent samples. The equivalence of item standard deviations is tested using the FMAX statistic, which is the ratio of the larger variance divided by the smaller variance, and which has an F-distribution. Since we are performing multiple tests of significance it is advisable to distribute the α-level (.05) equally over all 25-items; thus, it is advisable to consider a difference or FMAX value to be significant only if p<.002 = .05/25.

In addition to comparing means and standard deviations for the two forms ofany item, when we use pseudo-classical scores, we can also compare the item reliabilities discussed in Chapter VI. These reliabilities are provided in Table 7-13.

We can summarize the critical information in Tables 7-12 and 7-13 in the following manner.

Item	Pretest Differences in:			Posttest Differences in:		
No.	Mn's	SD's	r's	Mn's	SD's	r's
2	x		x			
3					x	
7		x				
9		x				
11			x	x	x	
13	x					
14		x				
15		x				
21					x	x
22		x	x			
23		x	x			
24				x	x	

In the above table, an "x" appears only if p<.002, and the items listed are only those for which at least one pretest or posttest difference is significant at p<.002. Clearly there is some evidence that the two forms of some items are not equivalent, for either the pretest mode or the posttest mode or both modes. Note that if two items are equivalent when administered in the pretest mode, this does not guarantee that the items will be equivalent when administered in the posttest mode, and vice-versa.

Data for Identifying Items that may Require Revision

In Chpater V the author specified a procedure for identifying items that may require revision. The basic data (or summary statistics) and rules for this procedure are summarized in Table 5-1. The results of applying this procedure (with some modifications and additions) to the items discussed in this Chapter are indicated in Tables 7-14 to 7-17. The reader should note that for each of these Tables: (a) each item was scored using the pseudo-classical scoring procedure; (b) pretest and posttest item reliabilities are considered as data for decision-making, along with the data discussed in Chapter V; (c) the Theoretical Error Rate (TER) is 0.75 for all items, since all items have four alternatives; and (d) an "x" indicates that revision may be required on the basis of the indicated rule.

The reader will note from the title for each of the four tables that: (a) the data reported in Table 7-14 are for the 31 subjects who took Form A for both the Pre- and Posttest, (b) the data reported in Table 7-16 are for the 28 subjects who took Form B for both the Pre- and Posttest, and (c) for Tables 7-15 and 7-17 the sets of subjects who took the Pre- and Posttests are not the same, although there is a considerable degree of overlap.

It should be noted that the decision rules specified in Tables 7-14 to 7-17 are, in several cases, based upon the author's subjective judgments with regard to the context within which the items were used. For example, there is no "objective" basis for saying that an item may need revision if PMPG<.50 -- others might argue for a cut-off value of, say, 0.40 or 0.60. It is also possible that another evaluator examining the same items might choose to add other statistics and/or decision rules, or an evaluator might even choose to eliminate certain statistics and/or rules. The important issues are that: (a) the decision rules be specified prior to an examination of the data, (b) the actual rules and cut-offs chosen have at least a logical basis for being stated, and (c) the procedure used for examining item data be systematic and, as much as possible, replicable.

A cursory analysis of Tables 7-14 to 7-17 will convince the reader that PMPG is often less than 0.50 and PER is often greater than 0.40. Thus, at a minimum, the instruction for the information tested by many of these criterion-referenced items has not been as effective as the author had hoped.

The actual task of determining which items to revise and what kinds of revisions to make involves: (a) using the item statistics and tests discussed in the previous section of this Chapter in order to determine those pairs of items that do not appear to be equivalent and (b) using statistics and rules of the kind reported in Tables 7-14 to 7-17 (as well as other supporting data such as item analysis tables) in order to determine which particular items and what aspects of such items require revision.

At the risk of being repetitious, we wish to state again that even if the data indicate that revision may be required, one must study the item carefully to determine what, if anything, needs revision. For example, there appear to be problems with both forms of item 21; yet, after analyzing the data, the item analysis tables for the two forms of the item, and the actual items themselves, no obvious problem with either item was apparent. Therefore, the author intends to retest both forms of item 21 at some future time, and if the same situation still prevails, then the author will eliminate or completely rewrite both items.

TABLE 7-1a

Means and Standard Deviations -- Pretest and Posttest
VAR(1) = Arithmetic Mean of Item Confidence Scores

	Pretest		Posttest		
	Fm A c_1d_1	Fm B c_2d_1	Fm A c_1d_2	Fm B c_2d_2	N
a_1b_1	.312 .060		.555 .142		21
a_2b_1	.373 .070		.602 .092		10
a_1b_2	.332 .046			.499 .150	19
a_2b_2	.342 .037			.516 .116	9
a_1b_3		.330 .049	535 .109		17
a_2b_3		.332 .064	.545 .119		9
a_1b_4		.322 .066		.493 .166	20
a_2b_4		.370 .064		.625 .141	8
$a_{\cdot}b_{\cdot}$	.333 .057	.333 .061	.556 .120	.518 .153	113

TABLE 7-1b

Means and Standard Deviations -- Pretest

VAR(1) = Arithmetic Mean of Item Confidence Scores

	Fm A	Fm A	Fm B	Fm B	Both
	b_1	b_2	b_3	b_4	$b_\bullet$
a_1	.312 .060 N=21	.332 .046 N=19	.330 .049 N=17	.322 .066 N=20	.324 .056 N=77
a_2	.373 .070 N=10	.342 .037 N=9	.332 .064 N=9	.370 .064 N=8	.354 .060 N=36
$a_\bullet$	.332 .068 N=31	.335 .043 N=28	.331 .053 N=26	.336 .068 N=28	.333 .059 N=113

TABLE 7-1c

Means and Standard Deviations -- Posttest

VAR(1) = Arithmetic Mean of Item Confidence Scores

	Fm A	Fm B	Fm A	Fm B	Both
	b_1	b_2	b_3	b_4	$b_\bullet$
a_1	.555 .142 N=21	.499 .150 N=19	.535 .109 N=17	.493 .166 N=20	.521 .144 N=77
a_2	.602 .092 N=10	.516 .116 N=9	.545 .119 N=9	.625 .141 N=8	.571 .120 N=36
$a_\bullet$	.570 .128 N=31	.505 .138 N=28	.538 .110 N=26	.530 .168 N=28	.537 .138 N=113

TABLE 7-2a

Means and Standard Deviations -- Pretest and Posttest

VAR(2) = Geometric Mean of Item Confidence Scores

	Pretest		Posttest		
	Fm A	Fm B	Fm A	Fm B	N
	c_1d_1	c_2d_1	c_1d_2	c_2d_2	
a_1b_1	.266 .040		.393 .119		21
a_2b_1	.251 .061		.375 .070		10
a_1b_2	.248 .049			.374 .139	19
a_2b_2	.239 .039			.344 .102	9
a_1b_3		.236 .034	.385 .101		17
a_2b_3		.267 .049	.362 .133		9
a_1b_4		.247 .045		.382 .136	20
a_2b_4		.289 .032		.478 .123	8
$a_{\bullet}b_{\bullet}$	.253 .047	.253 .044	.383 .107	.387 .133	113

TABLE 7-2b

Means and Standard Deviations -- Pretest

VAR(2) = Geometric Mean of Item Confidence Scores

	Fm A	Fm A	Fm B	Fm B	Both
	b_I	b_2	b_3	b_4	$b_\bullet$
a_1	.266	.248	.236	.247	.250
	.040	.049	.034	.045	.043
	N=21	N=19	N=17	N=20	N=77
a_2	.251	.239	.267	.289	.260
	.061	.039	.049	.032	.049
	N=10	N=9	N=9	N=8	N=36
$a_\bullet$	.261	.245	.247	.259	.253
	.048	.045	.042	.045	.045
	N=31	N=28	N=26	N=28	N=113

TABLE 7-2c

Means and Standard Deviations -- Posttest

VAR(2) = Geometric Mean of Item Confidence Scores

	Fm A	Fm B	Fm A	Fm B	Both
	b_1	b_2	b_3	b_4	$b_\bullet$
a_1	.393	.374	.385	.382	.384
	.119	.139	.101	.136	.123
	N=21	N=19	N=17	N=20	N=77
a_2	.375	.344	.362	.478	.387
	.070	.102	.133	.123	.115
	N=10	N=9	N=9	N=8	N=36
$a_\bullet$	.387	.364	.377	.410	.385
	.105	.127	.111	.137	.120
	N=31	N=28	N=26	N=28	N=113

TABLE 7-3a

Means and Standard Deviations -- Pretest and Posttest
VAR(3) = Arithmetic Mean of Item Log Scores

	Pretest		Posttest		
	Fm A	Fm B	Fm A	Fm B	N
	c_1d_1	c_2d_1	c_1d_2	c_2d_2	
a_1b_1	70.936 3.727		78.687 7.124		21
a_2b_1	68.233 6.254		78.402 3.929		10
a_1b_2	69.263 4.792			77.245 7.504	19
a_2b_2	68.629 3.626			75.892 7.019	9
a_1b_3		68.430 3.184	78.578 5.650		17
a_2b_3		70.928 4.298	76.836 6.965		9
a_1b_4		69.315 4.137		77.953 7.154	20
a_2b_4		72.930 2.413		83.365 5.395	
$a_{\cdot}b_{\cdot}$	69.757 4.543	69.841 3.892	78.312 6.090	78.189 7.213	113

TABLE 7-3b

Means and Standard Deviations -- Pretest

VAR(3) = Arithmetic Mean of Item Log Scores

	Fm A b_1	Fm A b_2	Fm B b_3	Fm B b_4	Both $b_{\cdot}$
a_1	70.936 3.727 N=21	69.263 4.792 N=19	68.430 3.184 N=17	69.315 4.137 N=20	69.549 4.044 N=77
a_2	69.233 6.254 N=10	68.629 3.626 N=9	70.928 4.298 N=9	72.930 2.413 N=8	70.327 4.601 N=36
$a_{\cdot}$	70.387 4.653 N=31	69.059 4.393 N=28	69.295 3.724 N=26	70.348 4.040 N=28	69.797 4.226 N=113

TABLE 7-3c

Means and Standard Deviations -- Posttest

VAR(3) = Arithmetic Mean of Item Log Scores

	Fm A b_1	Fm B b_2	Fm A b_3	Fm B b_4	Both $b_{\cdot}$
a_1	78.687 7.124 N=21	77.345 7.504 N=19	78.578 5.650 N=17	77.953 7.154 N=20	78.141 6.820 N=77
a_2	78.402 3.929 N=10	75.892 7.019 N=9	76.836 6.965 N=9	83.365 7.054 N=8	78.486 6.326 N=36
$a_{\cdot}$	78.595 6.203 N=31	76.878 7.254 N=28	77.975 6.055 N=26	79.500 7.054 N=28	78.251 6.640 N=113

TABLE 7-4a

Means and Standard Deviations -- Pretest and Posttest

VAR(4) = Arithmetic Mean of Item Elimination Scores

	Pretest		Posttest		
	Fm A c_1d_1	Fm B c_2d_1	Fm A c_1d_2	Fm B c_2d_2	N
a_1b_1	.150 .124		.552 .200		21
a_2b_1	.269 .112		.624 .098		10
a_1b_2	.194 .102			.444 .213	19
a_2b_2	.212 .079			.458 .172	9
a_1b_3		.186 .091	.563 .158		17
a_2b_3		.196 .132	.526 .199		9
a_1b_4		.171 .106		.525 .200	20
a_2b_4		.243 .105		.648 .138	8
$a_\bullet b_\bullet$	.194 .114	.191 .106	.564 .172	.504 .201	113

TABLE 7-4b

Means and Standard Deviations -- Pretest

VAR(4) = Arithmetic Mean of Item Elimination Scores

	Fm A b_1	Fm A b_2	Fm B b_3	Fm B b_4	Both $b_\bullet$
a_1	.150 .124 N=21	.194 .102 N=19	.186 .091 N=17	.171 .106 N=20	.174 .106 N=77
a_2	.269 .112 N=10	.212 .079 N=9	.196 .132 N=9	.243 .105 N=8	.231 .108 N=36
$a_\bullet$	.189 .131 N=31	.200 .094 N=28	.189 .104 N=26	.192 .109 N=28	.192 .110 N=113

TABLE 7-4c

Means and Standard Deviations -- Posttest

VAR(4) = Arithmetic Mean of Item Elimination Scores

	Fm A b_1	Fm B b_2	Fm A b_3	Fm B b_4	Both $b_\bullet$
a_1	.552 .200 N=21	.444 .213 N=19	.563 .158 N=17	.525 .200 N=20	.521 .197 N=77
a_2	.624 .098 N=10	.458 .172 N=9	.526 .199 N=9	.648 .138 N=8	.563 .168 N=36
$a_\bullet$	.575 .175 N=31	.448 .198 N=28	.550 .170 N=26	.560 .190 N=28	.534 .188 N=113

TABLE 7-5a

Means and Standard Deviations -- Pretest and Posttest

VAR(5) = Arithmetic Mean of Item Pseudo-Classical Scores

	Pretest		Posttest		
	Fm A	Fm B	Fm A	Fm B	N
	c_1d_1	c_2d_1	c_1d_2	c_2d_2	
a_1b_1	.360 .086		.658 .132		21
a_2b_1	.424 .081		.700 .069		10
a_1b_2	.378 .071			.569 .156	19
a_2b_2	.403 .046			.573 .108	9
a_1b_3		.367 .071	.666 .115		17
a_2b_3		.386 .096	.634 .143		9
a_1b_4		.357 .078		.626 .150	20
a_2b_4		.427 .091		.737 .097	8
$a_{\bullet}b_{\bullet}$	.383 .077	.376 .082	.664 .118	.614 .148	113

TABLE 7-5b

Means and Standard Deviations -- Pretest
VAR(5) = Arithmetic Mean of Item Pseudo-Classical Scores

	Fm A	Fm A	Fm B	Fm B	Both
	b_1	b_2	b_3	b_4	$b_\bullet$
a_1	.360 .086 N=21	.378 .071 N=19	.367 .071 N=17	.357 .078 N=20	.365 .076 N=77
a_2	.424 .081 N=10	.403 .046 N=9	.386 .096 N=9	.427 .091 N=8	.410 .079 N=36
a	.381 .088 N=31	.386 .064 N=28	.374 .079 N=26	.377 .086 N=28	.380 .079 N=113

TABLE 7-5c

Means and Standard Deviations -- Posttest
VAR(5) = Arithmetic Mean of Item Pseudo-Classical Scores

	Fm A	Fm B	Fm A	Fm B	Both
	b_1	b_2	b_3	b_4	b
a_1	.658 .132 N=21	.569 .156 N=19	.666 .115 N=17	.626 .150 N=20	.630 .142 N=77
a_2	.700 .069 N=10	.573 .108 N=9	.634 .143 N=9	.737 .097 N=8	.660 .120 N=36
$a_\bullet$	.672 .116 N=31	.570 .140 N=28	.655 .123 N=26	.658 .144 N=28	.639 .136 N=113

TABLE 7-6a

Means and Standard Deviations -- Pretest and Posttest

VAR(6) = Arithmetic Mean of Classical Scores

	Pretest		Posttest		
	Fm A	Fm B	Fm A	Fm B	N
	c_1d_1	c_2d_1	c_1d_2	c_2d_2	
a_1b_1	.404 .089		.691 .118		21
a_2b_1	.464 .076		.700 .063		10
a_1b_2	.444 .080			.634 .146	19
a_2b_2	.418 .098			.578 .104	9
a_1b_3		.419 .108	.678 .122		17
a_2b_3		.449 .115	.662 .122		9
a_1b_4		.414 .090		.644 .135	20
a_2b_4		.430 .102		.760 .117	8
$a_\bullet b_\bullet$	.429 .087	.424 .100	.685 .110	.646 .139	113

TABLE 7-6b

Means and Standard Deviations -- Pretest

VAR(6) = Arithmetic Mean of Classical Scores

	Fm A	Fm A	Fm B	Fm B	Both
	b_1	b_2	b_3	b_4	$b_\bullet$
a_1	.404 .089 N=21	.444 .080 N=19	.419 .108 N=17	.414 .090 N=20	.420 .091 N=77
a_2	.464 .076 N=10	.418 .098 N=9	.449 .115 N=9	.430 .102 N=8	.441 .095 N=36
$a_\bullet$	.423 .089 N=31	.436 .085 N=28	.429 .109 N=26	.419 .092 N=28	.427 .093 N=113

TABLE 7-6c

Means and Standard Deviations -- Posttest

VAR(6) = Arithmetic Mean of Classical Scores

	Fm A	Fm B	Fm A	Fm B	Both
	b_1	b_2	b_3	b_4	$b_\bullet$
a_1	.691 .118 N=21	.634 .146 N=19	.678 .122 N=17	.644 .135 N=20	.662 .131 N=77
a_2	.700 .063 N=10	.578 .104 N=9	.662 .122 N=9	.760 .117 N=8	.673 .118 N=36
$a_\bullet$	.694 .103 N=31	.616 .135 N=28	.672 .120 N=26	.677 .139 N=28	.666 .126 N=113

TABLE 7-7

Test Reliabilities for Six Different Types of Scores

Pretest A (N=59)				Pretest B (N=54)			
	Mean	SD	r		Mean	SD	r
VAR(1)	.333	.057	.547	VAR(1)	.333	.061	.599
VAR(2)	.253	.047	.324	VAR(2)	.253	.044	****
VAR(3)	69.757	4.543	.318	VAR(3)	69.841	3.892	.049
VAR(4)	.194	.114	.430	VAR(4)	.191	.106	.229
VAR(5)	.383	.077	.371	VAR(5)	.376	.082	.395
VAR(6)	.429	.087	-.041	VAR(6)	.424	.100	.130

Posttest A (N=57)				Posttest B (N=56)			
	Mean	SD	r		Mean	SD	r
VAR(1)	.556	.120	.799	VAR(1)	.518	.153	.892
VAR(2)	.383	.107	.359	VAR(2)	.387	.133	.733
VAR(3)	78.312	6.090	.442	VAR(3)	78.189	7.213	.677
VAR(4)	.564	.172	.677	VAR(4)	.504	.201	.756
VAR(5)	.664	.118	.631	VAR(5)	.614	.148	.757
VAR(6)	.685	.110	.474	VAR(6)	.646	.139	.646

Note.--All reliability coefficients, except those for VAR(2), were calculated using Hoyt's analysis of variance technique. When a subject's score is the geometric mean of the subjective probabilities associated with the correct answers to items [VAR(2)] , one cannot employ Hoyt's technique for calculating reliability; therefore, for VAR(2), we report odd-even split-halves coefficients.

**** indicates that the coefficient could not be calculated.

Note.--One can calculate Livingston's reliability coefficient for any criterion score or cut-off value using the means, standard deviations, and reliability coefficients reported above.

TABLE 7-8

Unweighted Means Analysis of Variance and Contrasts for Subject Pretest Scores

Using VAR(5) = Arithmetic Mean of Item Pseudo-Classical Scores

Source	SS	df	MS	F	
A	0.037	1	0.037	6.167	($p<.05$)
B	0.000	3	0.000	0.000	
AB	0.002	3	0.001	0.167	
w.cell	0.643	105	0.006		

		Levels of A and B												
		a_1b_1	a_2b_1	a_1b_2	a_2b_2	a_1b_3	a_2b_3	a_1b_4	a_2b_4	Value[1]	SE[1]	Test Stat[1]	Orth[2] t	Bonf[3] t
Contrasts	CT_1	1	-1	0	0	0	0	0	0	-.064	.030	-2.129	*	NS
	CT_2	0	0	1	-1	0	0	0	0	-.025	.032	-0.789	NS	NS
	CT_3	0	0	0	0	1	-1	0	0	-.019	.032	-0.589	NS	NS
	CT_4	0	0	0	0	0	0	1	-1	-.070	.033	-2.138	*	NS
	CT_5	.50	-.50	.50	-.50	0	0	0	0	-.089	.044	-2.038	*	NS
	CT_6	0	0	0	0	.50	-.50	.50	-.50	-.089	.046	-1.936	NS	NS
	CT_7	.25	.25	.25	.25	-.25	-.25	-.25	-.25	.028	.063	0.442	NS	NS
	CT_8	.50	.50	-.50	-.50	0	0	0	0	.003	.044	0.069	NS	NS
	CT_9	0	0	0	0	.50	.50	-.50	-.50	-.031	.046	-0.674	NS	NS

[1] SE = standard error; Value/SE = Test Stat.

[2] * implies $p<.05$ which occurs when the test statistic is greater than 1.99 .

[3] * implies $p<.05$ which occurs when the test statistic is greater than 2.85 .

TABLE 7-9

Unwieghted Means Analysis of Variance and Contrasts for Subject Posttest Scores

Using VAR$^{(5)}$ = Arithmetic Mean of Item Pseudo-Classical Scores

Source	SS	df	MS	F	
A	0.037	1	0.037	2.175	
B	0.195	3	0.065	3.821	($p<.05$)
AB	0.061	3	0.020	1.176	
w.cell	1.786	105	0.017		

		Levels of A an` B												
		a_1b_1	a_2b_1	a_1b_2	a_2b_2	a_1b_3	a_2b_3	a_1b_4	a_2b_4	Value[1]	SE[1]	Test Stat[1]	Orth t[2]	Bonf t[3]
Contrasts	CT_1	1	-1	0	0	0	0	0	0	-.042	.050	-0.838	NS	NS
	CT_2	0	0	1	-1	0	0	0	0	-.004	.053	-0.076	NS	NS
	CT_3	0	0	0	0	1	-1	0	0	.032	.054	0.595	NS	NS
	CT_4	0	0	0	0	0	0	1	-1	-.113	.056	-2.071	*	NS
	CT_5	.50	-.50	0	0	.50	-.50	0	0	-.010	.073	-0.136	NS	NS
	CT_6	0	0	.50	-.50	0	0	.50	-.50	-.117	.076	-1.541	NS	NS
	CT_7	.25	.25	-.25	-.25	.25	.25	-.25	-.25	.155	.106	1.467	NS	NS
	CT_8	.50	.50	0	0	-.50	-.50	0	0	.058	.073	0.789	NS	NS
	CT_9	0	0	.50	.50	0	0	-.50	-.50	-.219	.076	-2.885	*	*

[1] SE = standard error; Value/SE = Test Stat.

[2] * implies $p<.05$ which occurs when the test statistic is greater than 1.99 .

[3] * implies $p<.05$ which occurs when the test statistic is greater than 2.85 .

TABLE 7-10

Item Means and Standard Deviations

Using Confidence Probabilities

	Pretest Means			Posttest Means			Pre Stan Dev's			Post Stan Dev's		
Item	Fm A	Fm B	Diff	Fm A	Fm B	Diff	Fm A	Fm B	FMAX	Fm A	Fm B	FMAX
1	.251	.279	-.028	.289	.311	-.022	.081	.134	2.79***	.255	.215	1.40
2	.418	.242	.176**	.545	.378	.167**	.291	.163	3.21***	.278	.280	1.01
3	.518	.451	.067	.652	.575	.077	.262	.261	1.01	.244	.307	1.59
4	.389	.374	.015	.619	.533	.084	.292	.232	1.58	.320	.272	1.39
5	.310	.303	.007	.400	.407	-.007	.200	.158	1.60	.277	.264	1.11
6	.427	.357	.070	.634	.541	.093	.298	.254	1.38	.292	.286	1.04
7	.229	.297	-.068*	.349	.337	.012	.148	.198	1.78*	.211	.210	1.01
8	.270	.286	-.016	.625	.620	.005	.107	.144	1.82*	.267	.300	1.26
9	.364	.292	.072	.697	.608	.089	.201	.251	1.56	.271	.312	1.32
10	.238	.301	-.063	.251	.426	-.175**	.246	.236	1.08	.341	.328	1.08
11	.433	.586	-.153*	.619	.783	-.164*	.235	.301	1.65	.383	.285	1.81
12	.252	.308	-.056	.717	.635	.082	.124	.181	2.14**	.304	.326	1.15
13	.222	.378	-.156*	.418	.498	-.080	.188	.252	1.78*	.360	.366	1.03
14	.246	.241	.005	.632	.430	.202**	.041	.061	2.24**	.350	.322	1.19
15	.261	.247	.014	.637	.578	.059	.056	.023	5.90***	.330	.321	1.06
16	.270	.310	-.040	.750	.675	.075	.171	.174	1.04	.260	.318	1.49
17	.284	.303	-.019	.669	.623	.046	.129	.170	1.73*	.262	.274	1.09
18	.377	.298	.079*	.674	.474	.200**	.204	.164	1.56	.289	.276	1.10
19	.274	.274	.000	.687	.664	.023	.156	.106	2.15**	.302	.320	1.12
20	.257	.268	-.011	.227	.264	-.037	.179	.199	1.24	.230	.227	1.02
21	.759	.686	.073	.866	.756	.110**	.260	.305	1.38	.170	.265	2.43***
22	.238	.297	-.059	.392	.450	-.058	.153	.214	1.96*	.277	.303	1.20
23	.288	.261	.027	.536	.460	.076	.197	.108	3.37**	.362	.311	1.35
24	.358	.260	.098*	.398	.236	.162**	.195	.181	1.17	.237	.191	1.53
25	.408	.436	-.028	.610	.678	-.068	.219	.211	1.08	.293	.301	1.05
N	59	54		57	56		59	54		57	56	

Note.--All tests are two-tailed.
* p<.05 ** p<.01 *** p<.002 = .05/25

TABLE 7-11

Item Means and Standard Deviations

Using Classical Scores

	Pretest Means			Posttest Means			Pre Stan Dev's			Post Stan Dev's		
Item	Fm A	Fm B	Diff	Fm A	Fm B	Diff	Fm A	Fm B	FMAX	Fm A	Fm B	FMAX
1	.237	.321	-.084	.228	.357	-.129	.429	.471	1.21	.423	.483	1.30
2	.678	.278	.400***	.772	.554	.218*	.471	.452	1.09	.423	.502	1.[illegible]0
3	.898	.704	.194**	.947	.732	.215***	.305	.461	2.29**	.225	.447	3.93***
4	.475	.547	-.072	.807	.786	.021	.504	.503	1.00	.398	.414	1.08
5	.559	.556	.003	.561	.536	.025	.501	.502	1.00	.501	.503	1.01
6	.586	.519	.070	.807	.714	.093	.497	.504	1.03	.398	.456	1.31
7	.136	.259	-.123	.456	.393	.063	.345	.442	1.64	.502	.493	1.04
8	.254	.240	.014	.825	.821	.004	.439	.432	1.04	.384	.386	1.01
9	.780	.389	.391***	.895	.786	.109	.418	.492	1.39	.310	.414	1.79
10	.271	.314	-.043	.298	.446	-.148	.448	.469	1.09	.462	.502	1.18
11	.593	.722	-.129	.702	.929	-.227***	.495	.452	1.20	.462	.260	3.15***
12	.153	.566	-.413***	.895	.875	.020	.363	.500	1.90*	.310	.334	1.16
13	.136	.519	-.383***	.404	.589	-.185*	.345	.504	2.13**	.495	.496	1.01
14	.237	.148	.089	.754	.429	.325***	.429	.359	1.43	.434	.499	1.32
15	.237	.185	.052	.667	.750	-.083	.429	.392	1.20	.476	.437	1.18
16	.271	.370	-.099	.912	.839	.073	.448	.487	1.18	.285	.371	1.69
17	.509	.407	.102	.930	.929	.001	.504	.496	1.03	.258	.260	1.02
18	.627	.444	.183	.842	.589	.253**	.488	.502	1.06	.368	.496	1.82*
19	.237	.352	-.115	.807	.786	.021	.429	.482	1.26	.398	.414	1.08
20	.271	.333	-.062	.211	.286	-.075	.448	.476	1.13	.411	.456	1.23
21	.932	.870	.062	1.000	.946	.054	.254	.339	1.79*	.000	.227	----
22	.237	.389	-.152	.474	.518	-.044	.429	.492	1.32	.504	.504	1.00
23	.356	.500	-.144	.597	.607	-.010	.483	.505	1.09	.495	.493	1.01
24	.509	.167	.342***	.456	.107	.349***	.504	.376	1.80*	.502	.312	2.59***
25	.559	.519	.040	.860	.857	.003	.501	.504	1.01	.350	.353	1.02
N	59	54		57	56		59	54		57	56	

Note.--All tests are two-tailed.
* p<.05 ** p<.01 *** p<.002 = .05/25

TABLE 7-12

Item Means and Standard Deviations

Using Pseudo-Classical Scores

	Pretest Means			Posttest Means			Pre Stan Dev's			Post Stan Dev's		
Item	Fm A	Fm B	Diff	Fm A	Fm B	Diff	Fm A	Fm B	FMAX	Fm A	Fm B	FMAX
1	.253	.323	-.070	.304	.385	-.081	.213	.312	2.14**	.389	.421	1.17
2	.514	.202	.312***	.716	.478	.238**	.438	.294	2.22**	.420	.460	1.20
3	.742	.574	.168*	.889	.707	.182**	.355	.387	1.19	.257	.406	2.51***
4	.448	.421	.027	.809	.729	.050	.414	.397	1.09	.382	.401	1.10
5	.357	.397	-.040	.490	.491	-.001	.321	.358	1.24	.453	.423	1.15
6	.465	.381	.084	.781	.689	.092	.375	.392	1.09	.338	.398	1.39
7	.184	.330	-.146**	.420	.374	.046	.207	.325	2.45***	.409	.375	1.19
8	.280	.299	-.019	.816	.763	.053	.242	.233	1.07	.326	.352	1.17
9	.530	.293	.237***	.851	.731	.120	.377	.343	1.21	.315	.399	1.60
10	.250	.335	-.085	.285	.478	-.193*	.377	.376	1.01	.439	.450	1.05
11	.540	.699	-.159*	.671	.899	-.228***	.358	.389	1.18	.451	.282	2.56***
12	.251	.343	-.092*	.851	.781	.070	.197	.263	1.78*	.313	.357	1.30
13	.213	.443	-.230***	.447	.549	-.102	.270	.375	1.92*	.481	.471	1.05
14	.236	.236	.000	.719	.470	.249**	.139	.090	2.41***	.409	.447	1.19
15	.274	.242	.032	.724	.708	.016	.142	.049	8.29***	.376	.387	1.06
16	.275	.370	-.095	.904	.768	.136*	.333	.358	1.16	.253	.371	2.14**
17	.346	.326	.020	.871	.827	.044	.266	.307	1.33	.290	.310	1.14
18	.504	.360	.144*	.809	.582	.227**	.394	.337	1.37	.358	.419	1.37
19	.257	.279	-.022	.798	.771	.027	.264	.241	1.19	.355	.392	1.22
20	.275	.335	-.060	.228	.281	-.053	.314	.366	1.36	.351	.373	1.12
21	.912	.787	.125*	.991	.915	.076*	.232	.341	2.17**	.066	.235	12.58***
22	.229	.363	-.134*	.436	.479	-.043	.233	.359	2.36***	.389	.409	1.11
23	.309	.269	.040	.588	.539	.049	.276	.167	2.72***	.439	.426	1.06
24	.424	.239	.185**	.444	.147	.297***	.341	.334	1.04	.410	.264	2.41***
25	.517	.546	-.029	.759	.814	-.055	.372	.387	1.08	.378	.322	1.38
N	59	54		57	56		59	54		57	56	

Note.--All tests are two-tailed.
* p<.05 ** p<.01 *** p<.002 = .05/25

TABLE 7-13

Item Reliabilities

Using Pseudo-Classical Scores

Item	Pretest			Posttest		
	Fm A	Fm B	Diff	Fm A	Fm B	Diff
1	.236	.434	-.198**	.703	.736	-.033
2	.755	.527	.228***	.852	.834	.018
3	.647	.602	.045	.658	.782	-.124*
4	.681	.636	.045	.928	.800	.128*
5	.441	.526	-.085	.807	.704	.103*
6	.556	.641	-.085	.657	.727	-.070
7	.281	.470	-.189**	.675	.590	.085
8	.286	.255	.031	.696	.674	.022
9	.561	.558	.003	.769	.796	-.027
10	.745	.624	.121*	.930	.798	.132*
11	.507	.707	-.200***	.906	.861	.045
12	.203	.302	-.099	.760	.733	.027
13	.428	.560	-.132*	.920	.881	.039
14	.105	.044	.061	.814	.789	.025
15	.100	.013	.087	.696	.712	-.016
16	.547	.541	.006	.725	.759	-.034
17	.307	.422	-.115*	.736	.660	.076
18	.610	.485	.125*	.815	.709	.106*
19	.359	.284	.075	.769	.856	-.087
20	.486	.591	-.105*	.688	.679	.011
21	.659	.682	-.023	.480	.698	-.218***
22	.302	.548	-.246***	.605	.659	-.054
23	.351	.139	.212***	.782	.718	.064
24	.468	.603	-.135*	.669	.546	.123*
25	.545	.594	-.049	.768	.673	.095*
N	59	54		57	56	

* Diff > .10
** Diff ≥ .15
*** Diff ≥ .20

TABLE 7-14

Item Statistics and Decision Rules Using

Pretest = Form A; Subjects = b_1; N = 31;

Posttest = Form A; Subjects = b_1; N = 31

	Statistics												
Item	BER	DER[1]	PMPG	PER	PDI[2]	RB[3]	RP[4]	DER >0.	PMPG <.50	PER >.4	PDI ≠0.	RP< 0.6	RP< RB
1	.76	-.01	.55	.34	.14	.18	.67						
2	.46	.29**	.35	.30	.41	.79	.86	x	x				
3	.28	.47**	.79	.06	.04	.70	.61	x					x
4	.55	.20**	.65	.19	-.06	.70	.89	x					
5	.59	.16	.20	.47	.23	.42	.78		x	x			
6	.48	.27**	.48	.25	.42	.67	.59	x	x			x	x
7	.85	-.10	.39	.52	.18	.32	.61		x	x			
8	.74	.01	.64	.27	.33	.20	.60						
9	.46	.29**	.70	.14	.33	.58	.76	x					
10	.79	-.04	.19	.64	.14	.65	.91		x	x			
11	.46	.29**	.28	.33	.16	.52	.96	x	x				
12	.77	-.02	.83	.13	.24	.16	.84						
13	.75	.00	.20	.60	.36	.41	.87		x	x			
14	.77	-.02	.69	.24	.53	.05	.85						
15	.75	.00	.65	.26	.36	.00	.70						
16	.79	-.04	.82	.14	.18	.45	.80						
17	.67	.08	.-5	.10	.13	.27	.79						
18	.47	.28**	.57	.20	.23	.61	.84	x					
19	.74	.01	.76	.18	.58	.39	.85						
20	.71	.04	-.14	.81	.20	.56	.60		x	x			
21	.09	.66**	.78	.02	.05	.77	.51	x				x	x
22	.73	.02	.32	.50	.69**	.40	.65		x	x	x		
23	.71	.04	.44	.40	.18	.30	.83		x				
24	.67	.08	.19	.54	.09	.41	.56		x	x			
25	.41	.34**	.34	.27	.31	.53	.78	x	x				

[1] ** p<=.01 when DER>=.19

[2] ** p<=.01 when |PDI|>=.59; n_u=9 and n_1=10

[3] RB = reliability of pretest item.

[4] RP = reliability of posttest item.

TABLE 7-15

Item Statistics and Decision Rules Using

Pretest = Form A; Subjects = b_1 and b_2; N = 59;

Posttest = Form A; Subjects = b_1 and b_3; N = 57

	Statistics												
Item	BER	DER[1]	PMPG	PER	PDI[2]	RB[3]	RP[4]	DER >0.	PMPG <.50	PER >.4	PDI ≠0.	RP< 0.6	RP< RB
1	.75	.00	.07	.70	.05	.24	.71		x	x			
2	.49	.26**	.43	.28	.27	.76	.86	x	x				
3	.26	.49**	.58	.11	.12	.65	.65	x					
4	.55	.20**	.65	.19	.13	.69	.94	x					
5	.64	.11	.20	.51	.32	.42	.82		x	x			
6	.53	.22**	.58	.22	.29	.56	.66	x					
7	.82	-.07	.29	.58	.24	.28	.68		x	x			
8	.72	.03	.75	.18	.22	.29	.69						
9	.47	.28**	.68	.15	.33	.57	.78	x					
10	.75	.00	.05	.71	.38	.75	.94		x	x			
11	.46	.29**	.28	.33	.29	.51	.92	x	x				
12	.75	.00	.80	.15	.27	.20	.77						
13	.79	-.04	.30	.55	.21	.43	.93		x	x			
14	.76	-.01	.63	.28	.46**	.10	.82				x		
15	.73	.02	.62	.28	.38	.10	.70						
16	.72	.03	.86	.10	.13	.55	.73						
17	.65	.10	.80	.13	.23	.31	.74						
18	.50	.25**	.62	.19	.22	.62	.82	x					
19	.74	.01	.73	.20	.44	.36	.78						
20	.72	.03	-.07	.77	.31	.49	.69		x	x			
21	.09	.66**	.89	.01	.02	.67	.48	x				x	x
22	.77	-.02	.27	.56	.55**	.30	.61		x	x	x		
23	.69	.06	.41	.41	.25	.35	.79		x	x			
24	.58	.17	.03	.56	.25	.47	.68		x	x			
25	.48	.27**	.50	.24	.23	.55	.78	x					

[1] ** $p<=.01$ when DER>=.19

[2] ** $p<=.01$ when |PDI|>=.45; $n_u=14$ and $n_1=22$

[3] RB = reliability of pretest item.

[4] RP = reliability of posttest item.

TABLE 7-16

Item Statistics and Decision Rules Using

Pretest = Form B; Subjects = b_4; N = 28;

Posttest = Form B; Subjects = b_4; N = 28

	Statistics												
Item	BER	DER[1]	PMPG	PER	PDI[2]	RB[3]	RP[4]	DER >0.	PMPG <.50	PER >.4	PDI ≠0.	RP< 0.6	RP< RB
1	.68	.07	.22	.53	.06	.54	.82		x	x			
2	.83	-.08	.28	.60	-.07	.48	.93		x	x			
3	.46	.29**	.48	.24	.38	.49	.78	x	x				
4	.63	.12	.71	.18	.30	.72	.86						
5	.59	.16	.18	.48	.37	.48	.80		x	x			
6	.63	.12	.60	.25	.30	.62	.81						
7	.67	.08	.25	.50	.52	.45	.79		x	x			
8	.76	-.01	.73	.20	.30	.16	.71						
9	.67	.08	.58	.28	.41	.66	.91						
10	.63	.12	.21	.50	.01	.66	.85		x	x			
11	.26	.49**	.65	.09	.11	.83	.92	x					
12	.70	.05	.90	.07	.14	.20	.69						
13	.53	.22**	.30	.37	.59**	.51	.96	x	x		x		
14	.77	-.02	.43	.44	.80**	.04	.95		x	x	x		
15	.76	-.01	.71	.22	.39	.01	.77						
16	.61	.14	.72	.17	.30	.55	.81						
17	.71	.04	.73	.19	.25	.36	.75						
18	.60	.15	.30	.42	.22	.60	.75		x	x			
19	.73	.02	.79	.15	.22	.30	.84						
20	.66	.09	.05	.63	.33	.51	.82		x	x			
21	.24	.51**	.83	.04	.07	.69	.49	x				x	x
22	.62	.13	.23	.48	.38	.56	.75		x	x			
23	.77	-.02	.34	.51	.78**	.02	.81		x	x	x		
24	.74	.01	-.23	.91	.12	.67	.69		x	x			
25	.31	.44**	.61	.12	.21	.60	.57	x				x	x

[1] ** p<=.01 when DER>=.20

[2] ** p<=.01 when |PDI|>=.59; n_u=8 and n_l=14

[3] RB = reliability of pretest item.

[4] RP = reliability of posttest item.

TABLE 7-17

Item Statistics and Decision Rules Using

Pretest = Form B; Subjects = b_3 and b_4; N = 54;

Posttest = Form B; Subjects = b_2 and b_4; N = 56

	Statistics							DER	PMPG	PER	PDI	RP<	RP<
Item	BER	DER[1]	PMPG	PER	PDI[2]	RB[3]	RP[4]	>0.	<.50	>.4	≠0.	0.6	RB
1	.68	.07	.10	.61	.21	.44	.74		x	x			
2	.80	-.05	.35	.52	.02	.53	.84		x	x			
3	.43	.32**	.33	.29	.42	.61	.79	x	x				
4	.58	.17	.53	.27	.28	.64	.81						
5	.60	.15	.15	.51	.30	.53	.71		x	x			
6	.62	.13	.50	.31	.37	.65	.73						
7	.67	.08	.06	.63	.46	.47	.60		x	x			
8	.70	.05	.66	.24	.35	.25	.68						
9	.71	.04	.62	.27	.32	.56	.80						
10	.66	.09	.21	.52	.20	.63	.81		x	x			
11	.30	.45**	.67	.10	.12	.71	.87	x					
12	.66	.09	.67	.22	.32	.30	.74						
13	.56	.19**	.20	.45	.62**	.56	.89	x	x	x	x		
14	.76	-.01	.30	.53	.76**	.04	.80		x	x	x		
15	.76	-.01	.62	.29	.39	.01	.72						
16	.63	.12	.63	.23	.29	.54	.77						
17	.67	.08	.75	.17	.23	.42	.67						
18	.64	.11	.34	.42	.31	.49	.72		x	x			
19	.72	.03	.83	.23	.32	.28	.87						
20	.66	.09	-.09	.72	.36	.60	.68		x	x			
21	.21	.54**	.62	.08	.13	.69	.71	x					
22	.64	.11	.19	.52	.49	.55	.67		x	x			
23	.73	.02	.40	.46	.63**	.14	.73		x	x	x		
24	.76	-.01	-.12	.85	.04	.61	.55		x	x		x	x
25	.45	.30**	.58	.19	.26	.60	.68	x					

[1] ** p<=.01 when DER>=.19 [2] ** p<=.01 when |PDI|>=.50; n_u=10 and n_1=32

[3] RB = reliability of pretest item. [4] RP = reliability of posttest item.

CHAPTER VIII

Summary and Suggestions

Summary

Since this report is quite long (much longer, in fact, than the author had intended) and, in many cases, quite detailed, it seems advisable to provide the reader with a brief summary of each of the chapters.

Chapter I. The major purposes of this chapter are to provide a context within which this report fits, and to introduce the reader to distinctions in terminology. We indicate, for example, that distinctions can be made between criterion-referenced testing and mastery testing, in that mastery testing can be viewed as a specific kind of criterion-referenced testing. However, this distinction is not maintained very well in the literature; thus, in order to avoid confusion with previous literature, we have, in general, reserved the term "mastery" for those issues, statistics, etc. that have previously carried the label "mastery."

Chapter II. The major purpose of this chapter is to examine the relevance of classical test theory to criterion-referenced and mastery testing. We find that the classical test theory assumptions are general enough to form a basis for criterion-referenced and mastery testing; however, we question whether or not these assumptions are sufficient. Furthermore, we find that the binomial error model is more likely to be appropriate for most criterion-referenced tests than is the normal error model.

Chapter III. In this chapter, which is primarily a review of the literature, we consider the concept of validity with respect to criterion-referenced and mastery testing. We find that content validity is of paramount concern for criterion-referenced and mastery testing, since there is seldom available any extra-test criterion measure. Consequently, the validity of a criterion-referenced or mastery test is, from a practical point of view, very clearly tied to the procedur: whereby the test is developed. We find that the "i:em forms" procedure is highly desirable in that this procedure guarantees a certain degree of "objective-item congruence."

Chapter IV. In this chapter, which is primarily a review of the literature, we consider the concept of reliability with respect to criterion-referenced and mastery testing. Reliability issues are probably the most frequently discussed quantitative issues surrounding criterion-referenced and mastery testing. We report, criticize, and compare each of the major reliability indices that have been proposed in the literature, and we find that there is considerable disagreement (or, perhaps, confusion) among researchers with respect to reliability issues. In particular, researchers have often failed to distinguish between (a) reliability indices for criterion-referenced and mastery tests, (b) reliability in the sense of stability, equivalence, or internal consistency, and (c) reliability for measures of state and measures of change. There is also some evidence for confusion between indices for test reliability and indices for instructional effectiveness.

Chapter V. In general, the first four chapters treat criterion-referenced and mastery testing without directly considering issues that are specific to an analysis of individual items. Chapters V, VI, and VII, on the other hand, are primarily concerned with the analysis of criterion-referenced and mastery items, per se. In Chapter V we discuss statistics that have been suggested for analyzing such items, and we present a procedure for identifying items that may require revision. The procedure discussed necessitates calculating a set of statistics for each item and defining a set of rules to specify how to employ the item statistics in order to identify items that appear to require revision.

Chapter VI. For the most part, Chapter V involves the explicit assumption that items are scored in the classical correct/wrong manner. In Chapter VI we consider alternatives to the classical procedure. Specifically, we consider elimination scoring and various scoring procedures that entail the collection of subjective probabilities form each student for each alternative of an item. We find that elimination scoring is of questionable value in the analysis of criterion-referenced and mastery items, but scoring procedures that entail subjective probabilities appear to have promise. In particular, we define and examine a new kind of score called a "pseudo-classical score" which appears to be quite useful as a basis for examining the reliability and validity of criterion-referenced and mastery items.

Chapter VII. In this chapter we present a statistical analysis of a set of item data which we use to illustrate many of the statistics and procedures discussed in Chapters V and VI.

Appendix A. In this appendix we present the manual for DEC-TEST, a Fortran IV computer program written by the author. DEC-TEST uses subjective probabilities in order to calculate a number of student scores over items (typically associated with confidence testing or admissible probability measurement) and a number of item scores (including confidence, elimination, and pseudo-classical scores). Also, DEC-TEST has an extensive capability for item analysis. The manual in Appendix A provides and extensive guide to the use of DEC-TEST, a detailed explanation of all outputs, scores, and statistics, and an introduction to the use of subjective probabilities in testing.

Suggestions for the Researcher

It is probably safe to say that there are no definitive answers to any issue in criterion-referenced or mastery testing; thus, in a sense, every issue is a potential topic for research. However, I would like to identify a few issues which I feel are critical or often overlooked:

(a) We need better statistical and non-statistical models for considering criterion-referenced and mastery testing -- models in which assumptions and criteria are stated clearly a~~n~~d unambiguously. For example, I believe that we need a test-theoretic model for criterion-referenced testing that incorporates both random error and systematic error and that employs a definition of true scсre which is different from the classical definition.

(b) We need more integrated theoretical and practical work concerning the reliability and validity of criterion-referenced and mastery tests.

(c) We need alternative procedures for item construction. In particular, we need a better capability of constructing item forms for disciplines other than mathematics and the physical sciences.

(d) We need much more consideration of alternative procedures for scoring items and defining criterion performance. At the present time, almost exclusively, items are scored in a correct/wrong (1,0) manner and criterion performance is defined in terms of number of items correct. This implies that we are only concerned about whether or not a student can recognize or recall a correct answer, and we are not concerned about things like the degree of certainty that a student associates with his or her response. At any rate, it is difficult to believe that the classical correct/wrong procedure is the best, or the only appropriate, method for scoring items and defining criterion performance.

e) We need more consideration of issues surrounding the identification of inadequate criterion-referenced and mastery test items and procedures for revising such items.

Suggestions for the Practitioner

Chapter II, Chapter VI, parts of Chapter VII, and Appendix A are probably of marginal concern at the present time for most practitioners. However, the author feels that most practitioners should be familiar with the issues treated in the remaining parts of this report. In particular, attention should be given to Chapters I, III, and V. Also, the bibliography provided on the next few pages should be especially useful to most practitioners. The issue of reliability treated in Chapter IV is exceedingly important; however, it is unfortunately true that there is no gererally accepted procedure for calculating the reliability of a criterion-referenced or mastery test. Thus, the author suggests that practitioners study Chapter IV but be very cautious in using or interpreting any single index of reliability.

BIBLIOGRAPHY

Berger, R.J. A measure of reliability for criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Eduation, Minneapolis, 1970.

Block, J.H. *Mastery learning: Theory and practice.* New York: Holt, 1971.

Block, J.H. Mastery learning in the classroom: an overview of recent research. University of California, Santa Barbara, 1973.

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

Bloom, B.S., Hastings, J.T., & Madaus, G.F. *Handbook of formative and summative evaluation.* New York: McGraw, 1971.

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

Brennan, R.L. *Some statistical problems in the evaluation of self-instructional programs.* (Doctoral dissertation, Harvard University) Ann Arbor, Michigan: University Microfilms, 1970. No. 70-23080

Brennan, R.L. A generalized upper-lower item discrimination index. *Educational and Psychological Measurement*, 1972, *32*, 289-303.

Brennan, R.L. A model for the use of achievement data and time data in an instructional system. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1973. (a)

Brennan, R.L. Computer-assisted achievement testing in instruction. *Journal of Educational Technology Systems*, 1973, 2(1), 3-16. Also in Rollett, H.B. & Klaus, W. (Eds.) *Fortschritte und Ergebnisse der Bildunstechnologie*, West Germany: Ehrenwirth, 1973, 331-343. (b)

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

Brennan, R.L., & Stolurow,L.M. An empirical decision process for formative evaluation. Paper presented at the annual meeting of the American Educational Research Association, New York, February, 1971. (ERIC, ED 048 343)

Brown, F.G. *Principles of educational and psychological testing*. Hinsdale, Illinois: The Dryden Press, 1970.

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

Carroll, J.B. Importance of the time factor in learning. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1973.

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

Coombs, C.H., Milholland, J.E., & Womer, F.B. The assessment of partial knowledge. *Educational and Psychological Measurement*, 1956, *16*, 13-37.

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

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

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

Cronbach, L.J., & Furby, L. How should we measure change -- or should be? *Psychological Bulletin*, 1970, *74*, 68-80.

Cronbach, L.J., & Gleser, G.C. *Psychological tests and personnel decisions*. (2nd ed.) Urbana, Illinois: University of Illinois Press, 1965.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam,N. *The dependability of behavioral measurements*. New York: Wiley, 1972.

Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Dahl, T. Toward an evaluative methodology for criterion-referenced measures: objective item congruence. Paper presented at the annual meeting of the California Educational Research Association, San Diego, April, 1971.

de Finetti, B. Methods of discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 1965, 13, 87-123.

DuBois, P.H. The design of correlational studies in training. In R. Glaser (Ed.), Training research and education. Pittsburgh: University of Pittsburgh Press, 1962. Pp. 63-86.

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

Ebel, R.L. Some limitations of criterion-referenced measurement. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, March, 1970.

Echternacht, G.T. The use of confidence testing in objective tests. Review of Educational Research, 1972, 42, 217-236.

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

Ferguson, R.L. The development, implementation and evaluation of computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh, 1969.

Ferguson, R.L. Computer assistance for individualizing measurement. University of Pittsburgh, Learning, Research and Development Center, March, 1971.

Flanagan, J. Units, scores, and norms. In E.F. Lindquist (Ed.), Educational measurement, Washington, D. C.: American Council on Education, 1951. Pp. 695-763.

Gardner, E.F. Normative standard scores. Educational and Psychological Measurement, 1962, 22, 7-14.

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

Glaser, R., & Klaus, D.J. Proficiency measurement: Assessing human performance. In R. Gagne (Ed.), Psychological principles in system development. New York:Holt, 1962. Pp. 421-427.

Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In E. Thorndike (Ed.), Educational measurement. Washington, D.C.: National Council of Education, 1971.

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

Hambleton, R.K. A review of testing and decision-making procedures for selected individualized instructional programs. American College Testing Technical Bulletin No. 15, Iowa City, August, 1973.

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

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

Harris, C.W. (Ed.) Problems in measuring change. Madison, Wisc.: University of Wisconsin Press, 1963.

Harris, C.W. An index of efficiency for fixed-length mastery tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972. (a)

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

Harris, C.W. Note on the variances and covariances of three error types. Journal of Educational Measurement, 1973, 10, 49-50.

Harris, M.L., & Stewart, D.M. Application of classical strategies to criterion-referenced test construction. A paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

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

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

Hively, W., Maxwell, G., Rabehl, G., Senison, D., & Lundin, S. *Domain-referenced curriculum evaluation: a theoretical handbook and a case study from the MINNEMAST project.* CSE Monograph Series in Evaluation, Volume I, Center for the Study of Evaluation, Unive-sity of California, Los Angeles, 1973.

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

Hoyt, C.J. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, *6*, 153-160.

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

Jenkins, W.L. Triserial r -- a neglected statistic. *Journal of Applied Psychology*, 1956, *40*, 63-64.

Keats, J.A., & Lord, F.M. Atheoretical derivation of the distribution of mental test scores. *Psychometrika*, 1962, *27*, 59-72.

Keats, J.A. Some generalizations of a theoretical distribution of mental test scores. *Psychometrika*, 1964, *29*, 215-231.

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

Kriewall, T.E., & Hirsch, E. The development and interpretation of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, 1969. (ERIC ED 042 815).

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

Light, R.J. Issues in the analysis of qualitative data. In R. Travers (Ed.), *Second handbook of research on teaching*. Chicago: Rand McNally, 1973. Pp. 318-381.

Linn, R.L., Rock, D.A., & Cleary, T.A. The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, 1969, 129-146.

Linn, R.L., Rock, D.A., & Cleary, T.A. Sequential testing for dichotomous decision. *Research Bulletin*, RB-70-31, May, 1970, Princeton: Educational Testing Service.

Livingston, S.A. A reply to Harris' "An interpretation of Livingston's reliability coefficient for criterion-referenced tests. *Journal of Educational Measurement*, 1972, *9*, 31. (a)

Livingston, S.A. A criterion-referenced application of classical test theory. *Journal of Educational Measurement*, 1972, 9, 13-26. (b)

Livingston, S.A. Reply to Shavelson, Block, and Ravitch's "Criterion-referenced testing: comments on reliability." *Journal of Educational Measurement*, 1972, *9*, 139. (c)

Lord, F.M. Inference about true scores from parallel test forms. *Educational and Psychological Measurement*, 1959, 19, 331-336. (a)

Lord, F.M. Problems in mental test theory arising from errors of measurement. *Journal of the American Statistical Association*, 1959, 54, 472-479. (b)

Lord, F.M. A strong true-score theory with applications. *Psychometrika*, 1965, 30, 239-270.

Lord, F.M. Some theory for tailored testing. In H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper and Row, 1970, 139-183.

Lord, F.M. An empirical study of the normality and independence of errors of measurement in test scores. Psychometrika, 1960, 25, 91-104.

Lord, F.M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971, 8, 147-151.

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

Marshall, J.L. Reliability indices for criterion-referenced tests: A study based on simulated data. Paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans, February, 1973.

Merwin, J.C., & Womer, F.B. Evaluation in assessing the progress of education to provide bases of public understanding and public policy. In R. Tyler (Ed.), Educational evaluation: new roles, new means. The sisty-eighth yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1969.

Millman, J. Passing scores and test lengths for domain-referenced measures. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1972.

Nitko, A.J. Some considerations when using a domain-referenced system of achievement tests in instructional situations. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, March, 1970.

Nitko, A.J. A model for criterion-referenced tests based on use. Paper presented at the annual meeting of the American Educational Research Association, New York, February, 1971.

Novick, M.R. The axioms and principle results of classical test theory. Journal of Mathematical Psychology, 1966, 3, 1-18.

Novick, M.R., Lewis, C., & Jackson, P.H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-46.

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

Ozenne, D.G. Toward an evaluative methodology for criterion-referenced measures: Test sensitivity. CSE Report No. 72. Los Angeles: Center for the Study of Evaluation, Graduate School of Education, University of California at Los Angeles, 1971.

Popham, W.J. Indices of adequacy for criterion-referenced test items. In W. Popham (Ed.), Criterion-referenced measurement: An introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971. Pp. 79-98.

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

Rovinelli, R., & Hambleton, R.K. Some procedures for the validation of criterion-referenced test items. Amherst, Mass.: Center for Educational Research, School of Education, University of Massachusetts, June, 1973.

Sabers, D.L., & Kania, J.G. Item precision in criterion-referenced measurement. A paper presented at the annual meeting of the National Council for Measurement in Education, Chicago, April, 1972.

Saupe, J.L. Selecting items to measure change. Journal of Educational Measurement, 1966, 3, 223-228.

Savage, L.J. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association, 1971, 66, 783-801.

Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Shavelson, R.J., Block, J.H., & Ravitch, M.M. Criterion-referenced testing: Comments on reliability. Journal of Educational Measurement, 1972, 9, 133-137.

Shuford, E.H., Albert, A., & Massengill, H. Admissible probability measurement procedures. Psychometrika, 1966, 31, 125-145.

Stufflebeam, D.L. The use of experimental design in educational evaluation. *Journal of Educational Measurement*, 1971, *8*, 267-274.

Tucker, L.R., Damarin, F., & Messick, S. A base-free measure of change. *Psychometrika*, 1966, *31*, 457-473.

Wald, A. *Sequential analysis*. New York: Wiley, 1947.

Wilks, S.S. Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *The Annals of Mathematical Statistics*, 1946, *17*, 257-281.

Manual

for

DEC-TEST: A FORTRAN IV Computer Program For Decision-Theoretic Test Scoring and the Analysis of Item Data

by

Robert L. Brennan
Department of Education
SUNY at Stony Brook
Stony Brook, N. Y. 11790

October 1, 1973

[1]The research reported herein was performed pursuant to Grant No. OEG-2-2-2B118 with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

Preface

Methods of Administering and Scoring a Test

Classical Testing. The classical method of administering and scoring a test item necessitates that a student indicate which alternative he or she believes is correct. If the student picks the correct alternative, then the student receives one point, otherwise, the student receives zero points. This very simple procedure forms the basis for much of classical test theory, and this procedure is quite useful for many purposes. However, this procedure clearly does not provide differential information about the relative attractiveness of each alternative for the student. One way to approximate such information is through elimination scoring; one way to actually accumulate such information is through decision-theoretic testing.

Elimination Testing. In elimination testing, the student indicates which alternatives he or she believes are incorrect. The student gets the highest possible item score (usually 1.0) when he or she eliminates all alternatives except the correct answer; the student gets the lowest possible item score (usually -1.0) when he or she eliminates only the correct answer. If neither one of these two extreme conditions prevail, then the student gets an intermediate score that is determined according to a specific scoring rule. Thus, elimination scoring provides some information about the relative attractiveness of each alternative; but, for example, if a student eliminates two alternatives, we do not know whether or not the student feels more uncertain about one alternative than about the other.

Decision-Theoretic Testing.[1] In decision-theoretic testing a student responds to a test item by providing reported (observed) probabilities for each alternative for the item, such that the reported (observed) probabilities sum to unity. Although there are a number of ways an item can be scored in a decision-theoretic testing framework, the scoring system employed by DEC-TEST is the logarithmic

[1] What we refer to as "decision-theoretic testing" has been called, among other things, "confidence testing," "valid confidence testing," and "admissible probability measurement." Echternacht (1972) and Savage (1971) provide reviews of relevant literature concerning this topic.

scoring system. This system, as described in detail by Shuford et al. (1966), has a number of useful properties. One such property is called the "reproducing" property, which implies that a student will maximize his or her expected score if and only if the student's reported (observed) probabilities are identical to his or her degree-of-belief (true) probabilities. For example, if a student's degree-of-belief (true) probabilities for a three-alternative item are 0.50, 0.25, and 0.25, respectively, then the student will maximize his or her expected score only if he or she responds with reported (observed) probabilities of 0.50, 0.25, and 0.25, respectively.

According to Savage (1971):

> Proper scoring rules hold forth promise as more sophisticated ways of administering multiple-choice tests in certain educational situations. The student is invited not merely to choose one [answer] (or possibly none) but to show in some way how his opinion is distributed over the [answers], subject to a proper scoring rule or a rough facsimile thereof.
>
> Though requiring more student time per item, these methods should result in more discrimination per item than ordinary multiple-choice tests, with a possible net gain. Also, they seem to open a wealth of opportunities for the educational experimenter.

Reasons for Programming DEC-TEST

One of the principal reasons why the author undertook to program DEC-TEST was to examine each of the three scoring systems discussed above, especially with respect to their differential usefulness for item analysis in both norm-refernced and criterion-referenced situations. In order to do this DEC-TEST accepts decision-theoretic test data and estimates how a student would respond under elimination testing and classical testing rules. DEC-TEST can then perform an item analysis for each item for each type of testing procedure.

Other resons that motivated the author to program DEC-TEST include: (a) a desire to provide the capability of obtaining a detailed analysis of student and item performance under decision-theoretic testing, (b) a desire to provide the capability of comparing estimates of reliability for the three types of testing procedures discussed above, and (c) a desire to provide the capability of examining a number of different issues concerning the use of decision-theoretic testing as a tool for measurement and evaluation.

Features of DEC-TEST

DEC-TEST is a computer program for Decision-Theoretic Testing and the Analysis of Item Data. DEC-TEST was programmed using the FORTRAN IV, Level G, compiler on the IBM-370/155 computer at the State University of New York at Stony Brook.

DEC-TEST can accept any one of five different kinds of input, and it can produce as many as forty four different outputs. Both students and items can be identified alphanumerically, student identifications can be sorted, missing data features are available, items can be weighted, and items can have 2-5 alternatives. Included in the different kinds of possible output are: (a) listings of control cards, input data, and observed probabilities, (b) 102 variables for each student (calculated, printed and/or punched), (c) sophisticated item analysis routines for decision-theoretic, elimination, and classical testing, (d) eight different rosters of student item scores (calculated, printed, and/or punched), and (e) eight different kinds of reliability analyses plus a summary of all reliability analyses.

We caution the user of DEC-TEST in that many of the scores calculated and outputs provided are of very recent origin and require further study before their usefulness and/or validity will have been demonstrated.

Using this Manual

This manual is not intended to provide a completely detailed description of decision-theoretic testing. Many statements are made without an associated proof, and many parts of this manual assume some familiarity with decision-theoretic testing, classical test theory, statistics, and/or intermediate algebra. This manual is intended to be technically accurate, but technical accuracy sometimes militates against simple explanations.

For the most part, knowledge of FORTRAN IV is not required for running DEC-TEST. An exception to this general rule occurs in the definition of object-time format statements (see Sections II and III). Also, some knowledge of the IBM Job Control Language (JCL) is required (see Section VI).

Terminology and notation with regard to decision-theoretic testing, at the present time, have not been standardized. For example, "true confidences" or "true probabilities," as used in this manual, have been called elsewhere "degree-of-belief probabilities," "state probabilities," "internalized probabilities," and "personal probabilities"; "observed probabilities" have been called elsewhere "reported probabilities" and "assigned probabilities." Whether or not the terminology and notation used here represents the "best" choice is open to question; however, it is the author's intention that the terminology and notation used in this manual be consistent. One slight inconsistency known to the author is that the word "student" is used interchangeably with "subject."

Acknowledgements

The author would like to give special thanks to DR. Emir Shuford, who is currently employed by the Rand Corporation. Dr. Shuford has spent a considerable amount of time discussing decision-theoretic testing with the author, and Dr. Shuford is responsible in whole or in part for much of the theoretical work that forms the basis for the student variables that are calculated by DEC-TEST.

The author would also like to express his gratitude to Dr. David McMullen of the State University of New York at Stony Brook. Dr. McMullen has provided the author with many constructive criticisms of DEC-TEST from a user's viewpoint. For the most part, these criticisms have been taken into account in the current version of DEC-TEST.

Any errors in this manual or in DEC-TEST itself are totally the fault of the author, and should not be attributed to either of the above-mentioned persons.

Sample Input and Output

Sample input and output can be obtained from the author by writing to him at the following address:

Dr. Robert L. Brennan
Department of Education
SUNY at Stony Brook
Stony Brook, New York 11790

The author will also provide a source deck upon request, at a fee to cover cost of punching deck, handling and shipping.

Table of Contents

Section Page

Section Page

List of Tables

List of Figures

I. Introduction to DEC-TEST and Decision-Theoretic Testing

In the following paragraphs of this section, we provide an introduction to the subject of decision-theoretic testing, which allows us to establish a notational scheme for subsequent sections. Also, we introduce the user to fundamental student test scores reported by DE?-TEST. Finally, we discuss different kinds of item scores based upon decision-theoretic scoring, elimination scoring, and classical scoring.

Section V of this manual may be considered as a continuation of Section I, in that Section V provides a discussion of, and formulas for, all of the 102 Individual Subject Scores reported by DEC-TEST. Thus, some users may find it beneficial to read Section V immediately after Section I.

Logarithmic Scoring System used by DEC-TEST

In decision-theoretic testing, the student assesses the "confidence" he or she has in the correctness of each of the alternatives for each item and expresses this "confidence" (directly or indirectly) in terms of probabilities. Let

P_{hij} = observed probability for student h (h = 1, 2, ..., N), for item i (i = 1, 2, ..., K), for alternative j (j = 1, 2, ..., n_i), where

N = number of students who took test,

K = number of items on test, and

n_i = number of alternatives for item i, (2 <= n_i <= 5).

Note that, in DEC-TEST, the number of alternatives for an item must be 2, 3, 4 or 5. Now, it can be shown that if a linear scoring system (e.g., sum of probabilities associated with correct answer or a linear function of this sum) were used, then it would be in a student's best interest to use probabilities of 1.0 and 0.0, only, regardless of the student's "true confidence" in each of the alternatives. By "best interest" we mean that the student would, in the long run, maximize his or her score, given his or her actual knowledge of the answers to the test items. Thus, a linear scoring system would motivate a student to guess.

In decision-theoretic testing we seek (as one of our goals) the elimination of guessing by defining a scoring system such that it is in the student's best interest to respond with probabilities that are isomorphic with the student's true confidence in each of the alternatives to every item. Shuford et al. (1966) have shown that, to fulfill these requirements, one can use a logarithmic scoring function defined as:

(1) $L_{hi} = A_i \log P_{hi*} + B_i$, where

L_{hi} = log score for student i,

log = $\log_{10}$,

A_i, B_i = parameters for log scoring function (discussed below and in Section III -- Third Input Card), and

P_{hi*} = observed probability associated with correct answer (j = *) on item i for student h.

Actually, L_{hi} has a lower limit equal to $-\infty$ when $P_{hi*} = 0.0$. Therefore, we truncate the function at a convenient point C_i, $0.0 < C_i < 1.0$, such that the lowest possible value for L_{hi} is

$$L_{hi} = A_i \log C_i + B_i \, .$$

Now, B_i is actually the highest possible value of L_{hi}, so the range of L_{hi} is

$$A_i \log C_i \, .$$

DEC-TEST allows the user to specify as many different sets of values for A_i, B_i, and C_i as there are different item types. An item type is defined as the number of alternatives an item has. Thus, for example, if a test is composed of two and three alternative items, then the number of different item types is two, and two (possibly different) sets of values for A_i, B_i, and C_i may be specified. However, in this section, for illustrative purposes, we will use $A_i = 50$, $B_i = 100$, and $C_i = 0.01$ for all items, i, regardless of the number of alternatives. The scoring function we will use is, thus,

$$L_{hi} = 50 \log P_{hi*} + 100$$

with a truncation value (Iowest acceptable value of P_{hi*}) of 0.01, and a range of 100 "points" for each item.

Consider the item parameters and observed probabilities for student h in Table A-1. Note that w_i is the weight for item i, and recall that * indicates the correct alternative. The log scores, L_{hi}, for each of the items are given in Table A-2. The weighted sum of these scores is

$$(2) \qquad L_{h+} = \sum_{i=1}^{K} w_i L_{hi}$$

$$= (1)(100) + (1)(35) + \cdots + (2)(95)$$

$$= 810,$$

where "+" indicates "sum." The weighted average of the log scores is

$$(3) \qquad L_{h.} = L_{h+} / \sum_{i=1}^{K} w_i$$

$$= 810/10$$

$$= 81,$$

where "." indicates "average."

Now, it can also be shown that:

$$(4) \qquad L_{h.} = A_i \log \{[\prod_{i=1}^{K} P_{hi*} \mathrm{EXP}(w_i)] \mathrm{EXP}(1/ \sum_{i=1}^{K} w_i)\} + B_i$$

where "EXP" means "exponential"; i.e., $L_{h.}$ is the log score that results when the geometric mean of the P_{hi*} (terms within braces in (4), above) replaces P_{hi*} in (1), above. For out illustrative data, the geometric mean is:

$$[(1.00)^1 (0.05)^1 \cdots (0.40)^2 (0.80)^2] \times$$

$$\mathrm{EXP}[1/(1 + 1 + \cdots + 2 + 2)]$$

TABLE A-1

Illustrative Data:

Observed and Adjusted Probabilities

Item Parameters				Obs. Probs.			Adj. Probs.		
i	n_i	w_i	*	p_{hi1}	p_{hi2}	p_{hi3}	$\hat{p}_{hi1}$	$\hat{p}_{hi2}$	$\hat{p}_{hi3}$
1	2	1	1	1.00	0.01		0.79	0.14	
2	2	1	1	0.05	0.95		0.17	0.76	
3	2	1	1	0.60	0.40		0.53	0.40	
4	2	1	1	0.45	0.55		0.43	0.50	
5	2	1	1	0.50	0.50		0.47	0.47	
6	3	1	1	0.20	0.40	0.40	0.27	0.40	0.40
7	3	2	1	0.40	0.30	0.30	0.40	0.33	0.33
8	3	2	1	0.80	0.20	0.01	0.66	0.27	0.14

TABLE A-2

Illustrative Data:

Unweighted Item Scores

i	n_i	w_i	*	P_{hi}	L_{hi}	$\hat{P}_{hi}$	$\hat{L}_{hi}$	E_{hi}	C_{hi}	I_{hi}	$\hat{I}_{hi}$	EN_{hi}	$\widehat{EN}_{hi}$
1	2	1	1	1.00	100	0.79	95	1.0	1.0	0.93	0.36	0.07	0.67
2	2	1	1	0.05	35	0.17	62	-1.0	0.0	0.71	0.29	0.29	0.74
3	2	1	1	0.60	89	0.53	86	1.0	1.0	0.03	0.01	0.97	1.02
4	2	1	1	0.45	83	0.43	82	0.0	0.0	0.01	0.00	0.99	1.03
5	2	1	1	0.50	85	0.47	84	0.0	0.5	0.00	0.00	1.00	1.03
6	3	1	1	0.20	65	0.27	73	-1.0	0.0	0.06	0.02	1.52	1.57
7	3	2	1	0.40	80	0.40	80	0.0	1.0	0.01	0.01	1.57	1.58
8	3	2	1	0.80	95	0.66	91	1.0	1.0	0.80	0.29	0.78	1.30

Note.--RETO, the tolerance for elimination scoring, is 0.10.

$= (0.00013824)\text{EXP}(0.1)$

$= 0.411$

$L_{h.}$ and the geometric mean (sometimes transformed linearly) are probably the two most common scores for decision-theoretic testing.

Clarification. In the foregoing discussion of the logarithmic scoring system, we assumed that a student responded with, what we call here, "observed probabilities." Actually, DEC-TEST allows the user to employ any one of five different kinds of input, which are first converted to "original probabilities," R_{hij}, and then converted to observed probabilities P_{hij}. These conversion procedures are discussed later in detail. Here we merely note that the task of converting input to original probabilities involves straightforward transformations, whereas the task of converting original to observed probabilities is one of resolving inconsistencies. A typical inconsistency results when the sum of the original probabilities does not equal unity. In this case, if the discrepancy is big enough (as defined by the user via the DCT parameter -- see Section III), the user can perform either one of two different types of normalization procedures (see NORM parameter in Section III) in order to produce the observed probabilities. The careful reader probably noted that for items numbered 1 and 8 in the illustrative data, the sum of the observed probabilities does not equal unity; however, the observed probabilities reported in Table A-1 are legitimate results given one type of normalization procedure available to the user of DEC-TEST.

Realism Line and Adjusted Probabilities

One obvious question when using decision-theoretic testing is, "To what extent is a student being realistic in the assignment of his or her probabilities?" If a student is totally realistic, then, for each of the probabilities he or she uses, the proportion of times each probability is correct will equal the probability itself. Graphically, as indicated by the solid line in Figure A-1, this implies that the Ideal line (meaning ideal realism) has a slope of 1.0 and an intercept of 0.0 . To the extent that this is not true, then the student is unrealistic, to some degree.

TABLE A-3

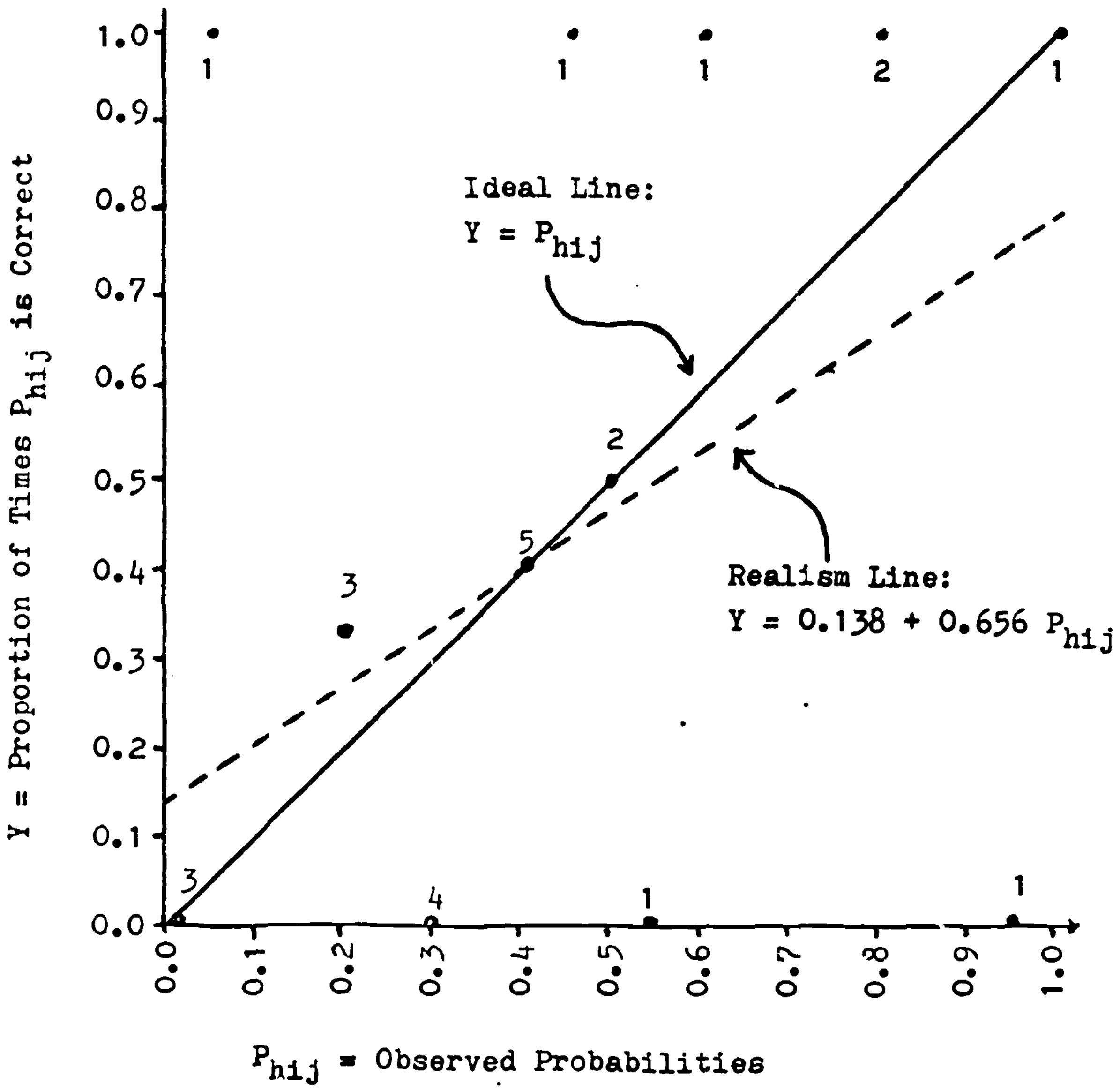
Illustrative Data: Proportion of Times that Distinct Observed Probability Values are Correct

Observed Probability P_{hij}	Weighted Number of Times Observed Probability is:		Proportion of Times P_{hij} Is Correct
	Used	Correct	
1.00	1	1	1.00
0.95	1	0	0.00
0.80	2	2	1.00
0.60	1	1	1.00
0.55	1	0	0.00
0.50	2	1	0.50
0.45	1	1	1.00
0.40	5	2	0.40
0.30	4	0	0.00
0.20	3	1	0.33
0.05	1	1	1.00
0.01	3	0	0.00
Totals	25	10	

FIGURE A-1

Illustrative Data

Ideal Line and Realism Line



Viewing our observed probabilities as indicated in Table A-3, we can plot the points in Figure A-1. Now, using least squares analysis to obtain a best-fitting straight line for these points, we obtain the Realism Line in Figure A-1, which has a slope of 0.656 and an intercept of 0.138 . (Formulas for the slope and intercept are provided in Section V.) Clearly, student h is somewhat unrealistic; in fact, student h is somewhat over-confident. (Some thought will convince the user that whenever the slope of the Realism Line is less than 1.0, the student is "over-confident." See Section V for other indicators of over- and under-confidence.)

Now, if the student had been more realistic, th student's observed probabilities would have been less extreme. For example, from the equation for the Realism Line, we note that, 0.138 + 0.656(0.80) = 0.66, which can be interpreted as meaning that the student would have been more realistic if he or she used 0.66 in place of 0.80. These new probabilities are called "adjusted probabilities" and denoted

(5) $\hat{P}_{hij} = \alpha_h + \beta_h P_{hij}$, where

α_h = intercept of Realism Line for student h, and

β_h = slope of Realism Line for student h.[1]

The set of adjusted probabilities for our illustrative data is provided in Table A-1 .

Now, using (5), above, it can be shown that:

$$(6) \quad \hat{P}_{hi+} = \sum_{j=1}^{n_i} \hat{P}_{hij}$$

$$= n_i \alpha_h + \beta_h \quad .$$

[1] We do, however, in practice impose two constraints on (5); namely, if

$\hat{P}_{hij} < C_i$, we set $\hat{P}_{hij} = C_i$, and if

$\hat{P}_{hij} > 1.0$, we set $\hat{P}_{hij} = 1.0$.

If all items on a test have the same number of alternatives, then the sum represented by (6), above, will equal 1.0 for all items and for all students who took the test. However, for our illustrative data we have both two- and three- alternative items; therefore, the sum of the adjusted probabilities for any given item is not 1.0 . In fact, for this data,

$$\hat{P}_{hi+} = 0.93 \text{ for items with two alternatives, and}$$

$$\hat{P}_{hi+} = 1.07 \text{ for items with three alternatives.}$$

Now, recall that the use of the log scoring function enables a student to maximize his or her score if the student is realistic. Since the adjusted probabilities are "more realistic" than the observed probabilities, it follows that, if

$$\hat{P}_{hi*} \text{ replaces } P_{hi*}$$

in (1), then, for a reasonably large number of items, $\hat{L}_{h.}$ (weighted mean of adjusted log scores) should exceed $L_{h.}$ (weighted mean of observed log scores). The $\hat{L}_{hi}$ scores are found in Table A-2, and their weighted mean $\hat{L}_{h.}$ is about 82. Thus, by being more realistic, student h could increase his or her average log score by approximately 82 - 81 = 1 "point."[1]

Taking the geometric mean of the $\hat{P}_{hi*}$ and using it in (1) gives approximately 0.44. Thus, the difference is geometric mean probability scores is approximately 0.44 - 0.41 = 0.03. Note that differences of 1 "point" and 0.03 are relatively small; yet, the difference in slopes between the Ideal Line and the Realism Line is reasonably large (0.344). This discrepancy demonstrates the need for caution in over-interpreting the meaning of differences between slopes, especially for very small amounts of data.

[1] As indicated previously, the sum of the adjusted probabilities for an item is not necessarily equal to 1.0 . Therefore, adjusted log scores, based upon adjusted probability scores, are somewhat biased when not all items in a test have the same number of alternatives. However, in the author's experience, any bias that exists in $\hat{L}_{h+}$ or $\hat{L}_{h.}$ is very slight, for most data.

Item Scores Computed by DEC-TEST

We have discussed above four different scores for item i for student h which are computed by DEC-TEST:

$P_{hi*} = P_{hi}$ = observed probability associated with correct answer,

$L_{hi*} = L_{hi}$ = observed log score (associated with correct answer),

$P_{hi*} = P_{hi}$ = adjusted probability associated with correct answer, and

$L_{hi*} = L_{hi}$ = adjusted log score (associated with correct answer).

Note that, throughout this manual, we use "^" to indicate that a variable makes use of adjusted probabilities.

We will now consider six other item scores generated by DEC-TEST.

Perceived Entropy and Perceived Information. One of the distinct advantages of decision-theoretic testing is that the availability of probabilities associated with each alternatove for an item allows us to interpret student responses in terms of information theoretic principles (see, for example, Shannon & Weaver, 1949). Thus, perceived entropy for item i for shudent h can be defined as:

$$(7) \qquad EN_{hi} = -\sum_{j=1}^{n_i} P_{hij} \log_2 P_{hij} .$$

Note that a good translation of "entropy" for our purposes is "uncertainty." Now, since the sum of observed probabilities for an item should equal unity, the maximum possible amount of perceived information is:

$$(8) \qquad MI_{hi} = \log_2 n_i ,$$

which implies that perceived information for item i is given by:

$$(9) \qquad I_{hi} = MI_{hi} - EN_{hi} .$$

Using Table A.1, the user can verify the values for perceived entropy and perceived information given in Table A 2. Note that $I_{hi} = 0.0$ when all observed probabilities equal $1/n_i$. (See, for example, Item No. 8.)

Actual Entropy and Actual Information. Using adjusted probabilities we define actual entropy in a manner analogous to that used to define perceived entropy; namely,

$$(10) \qquad \hat{EN}_{hi} = -\sum_{j=1}^{n_i} \hat{P}_{hij} \log_2 \hat{P}_{hij} .$$

Now, in order to define actual information, we need to know the maximum possible amount of actual information, which is, in general,

$$(11) \qquad \hat{MI}_{hi} = -(n_i\alpha_h + \beta_h)[\log_2(n_i\alpha_h + \beta_h) - \log_2(n_i)] .$$

In fact, (11) reduces to (8) when all items have the same number of alternatives, in which case the sum of the adjusted probabilities for any item equals unity.[1]

Using (10) and (11), actual information is defined as:

$$(12) \qquad \hat{I}_{hi} = \hat{MI}_{hi} - \hat{EN}_{hi} .$$

Using Table A-1, the user can verify the values for actual entropy and actual information given in Table A-2. Note that, for this data:

$\hat{MI}_{hi} = 1.027$ for items with two alternatives, and

$\hat{MI}_{hi} = 1.591$ for items with three alternatives.

Elimination Scores. Coombs et al. (1956) suggest a procedure for scoring a test based upon considering the alternatives that a student eliminates, i.e., judges to be incorrect. For this scoring system, a student receives $1/(n_i - 1)$ points for each incorrect alternative

[1] Even formula (11) is apt to be somewhat inaccurate in that we never allow an adjusted probability to be less that the truncation value for the log scoring function, and we never allow an adjusted probability to be greater than unity. However, any bias in (11) is usually very slight.

eliminated and loses 1.0 point for eliminating a correct alternative; thus, an unweighted item score falls between -1.0 and 1.0 .

Now, using the observed probabilities for item i, one can estimate a student's elimination score for item i, which we designate as E_{hi} .

There is, however, one problem in this estimation procedure, which can be illustrated by considering the observed probabilities for Item No. 4,

$$P_{h41} = 0.45 \text{ and } P_{h42} = 0.55 ,$$

in the illustrative data in Table A.1. If student h had taken this item under elimination scoring rules, would student h eliminate only alternative-1, both alternatives, or neither alternative? Actually, if student h eliminated both alternatives or neither alternative, the elimination score for the item would be the same, namely, 0.0 . However, whether or not student h will eliminate alternative-1 depends upon whether or not, for elimination scoring purposes, student h would consider probability differences of 0.55 - 0.45 = 0.10 meaningful and significant.

Thus, in order to estimate an elimination score from the observed probabilities for an item, the user of DEC-TEST must first assign a value to

RETO = tolerance for elimination scoring.

This single value for RETO is used by DEC-TEST to estimate elimination scores for all items, for all students. Letting PMAX be the magnitude of the largest observed probability for item i, for student h, E_{hi} is determined by applying the following algorithm to each of the n_i alternatives of item i for student h:

(a) If (P_{hij} + RETO - PMAX) >= 0.0 , add 0.0 ;

(b) If (P_{hij} + RETO - PMAX) < 0.0 and $j = *$, subtract 1.0 ; or

(c) If (P_{hij} + RETO - PMAX) < 0.0 and $j \neq *$, add $1/(n_i - 1)$.

Note that (b) and (c) indicate eliminated alternatives.

For example, if RETO = 0.10, then student h would not eliminate any of the probabilities for Item No. 4, resulting in $E_{h4} = 0.0$. See Table A-2 for other examples.

Classical Scores. For the classical scoring system, the student is forced to pick one and only one alternative. Using the observed probabilities, we can estimate a student's item score for the classical system. The procedure is as follows: if item i has ℓ ($\ell <= n_i$) highest probabilities, one of which is associated with the correct answer, then the classical unweighted item score for student h is

$$C_{hi} = 1/\ell \text{ ; otherwise,}$$

$$C_{hi} = 0.0 \text{ .}$$

For example, since, for Item No. 5, both observed probabilities are 0.50 (one of which is obviously associated with the correct answer , since the item has only two alternatives), $C_{h5} = 1/2 = 0.50$; i.e., if forced to pick only one alternative, student h has a 50-50 chance of picking the correct answer and getting 1.0 point. Note that $C_{h6} = 0.0$, since neither of the two highest probabilities is associated with the correct answer.

II. Summary of Control Cards and Student Data

The control cards and student data constitute the input to DEC-TEST. All control cards except the Item Type Key, Answer Key, and Item Cards must be on logical unit LUCC, which is specified in the main-line program of DEC-TEST. Usually, LUCC = 5, since most installations use logical unit 5 for reading punched 80-column cards (the usual medium for control cards) in FORTRAN; however, this value can be altered (see Section VI).

No knowledge of FORTRAN is necessary for setting up the control cards except for those control cards that are object-time formats.

For most of the control cards, the following information is provided: (a) data field (card columns), (b) variable identification, and (c) a brief description of the variable and its possible values.

Unless otherwise specified in the description of the variable, variables beginning with I, J, K, L, M, or N are integers, and other variables are real variables. The values of integer variables should be right-justified integers without any decimal point. The values of real variables should be (a) right justified integers or (b) decimal numbers including the decimal point. Technically, decimal numbers need not be right-justified, but it is usually desirable to right-justify them anyway. When A-format is specified (see RUN(5) and JT(i) in the First Input Card and the Item Cards, respectively), any alphabetic or numeric (alphanumeric) character may be used.

Note that, for most of the control cards, columns 1-10 are not used by DEC-TEST. For such control cards, these columns may be employed by the user for card identification. Often, descriptions of variables in the following cards end with "(0,1)". In these cases, the possible values for the variable are "0" meaning "no" or "absent" and "1" meaning "yes" or "present."

The user is cautioned that for integer and real variables, FORTRAN interprets blanks as "0" and "0.0", respectively.

First Input Card (required)

Columns	Variable	Description
11-15	RUN(5)	Five character run identification (A-format)
16-20	K	Number of items
21-25	N	Number of students (student records) 0 = until end of file
26-30	INC	No. of columns for student identification ($0 \leq$ INC ≤ 24) 0 = no student identifications
31-35	INCS	No. of columns for sort ($0 \leq$ INCS $\leq$ INC ≤ 24) 0 = no sort
36-40	IBGS	First column for sort (if INC $\neq 0$ and INCS $\neq 0$, $1 \leq$ IBGS $\leq$ INC ≤ 24 and IBGS + INCS - 1 $\leq$ INC)
41-45	IXTRA	Additional student variable (0,1)
46-55	XMS	Missing data code
56-60	MSD	Technique for handling missing data for an item 0 = alternatives for missing item transformed to observed probabilities of 1/no. of alternatives 1 = skip item
61-65	IOTH	Number of object-time format cards for Heading ($1 \leq$ IOTH ≤ 2)
66-70	IOTD	Number of object-time format cards for Student Data Input ($1 \leq$ IOTD ≤ 10)
71-75	INVAR	Number of student variables on Second Input Card(s) 0 = ten default student variables

(2) Second Input Card (required)

Columns	Variable	Description
11-15	IPT	Kind of student data used as input 1 = probabilities 2 = 100 X probabilities 3 = star method 4 = log scores 5 = log scores in which, for all items, $A_i = 50$, $B_i = 100$, $C_i = 0.01$, and scores of 100 are input as 99
16-25	DCT	Tolerance for decision-theoretic testing
26-30	NORM	Normalization procedure 0 = no normalization 1 = normalize over all probabilities for item, except those equal to truncation value 2 = normalize over all probabilities
56-60	LUCD	Logical unit for reading Item Cards or Item Keys
61-65	LUSD	Logical unit for reading Student Data Input
66-70	LUPT	Primary logical unit for printing
71-75	LUPT2	Secondary logical unit for printing
76-80	LUPC	Logical unit for punching

(3) Third Input Card (required)

Columns	Variable	Description	
11-15	AB(2,1)	Low score	two alternatives
16-20	AB(2,2)	High score	
21-27	AB(2,3)	Truncation value	
28-32	AB(3,1)	Low score	three alternatives
33-37	AB(3,2)	High score	
38-44	AB(3,3)	Truncation value	
45-49	AB(4,1)	Low score	four alternatives
50-54	AB(4,2)	High score	
55-61	AB(4,3)	Truncation value	
62-66	AB(5,1)	Low score	five alternatives
67-71	AB(5,2)	High score	
72-78	AB(5,3)	Truncation value	

Note: If scoring function to be used is such that log score = 0.0 when probability associated with correct answer is 1/no. of alternatives, then AB(n_i,3) may be left blank and will be automatically calculated by DEC-TEST.

(4) Item Analysis Definition Card -- Decision-Theoretic Scoring (required)

Columns	Variable	Description		
11-15	IDCV	Criterion variable number		
16-20	IDGP	Groupting parameter 1 = percent of subjects 2 = score range		
21-30	RDL(1)	LIMIT(1)	Low group	
31-40	RDL(2)	LIMIT(2)		Middle gp
41-50	RDL(3)	LIMIT(3)	High gp	
51-60	RDL(4)	LIMIT(4)		

(5) Item Analysis Definition Card -- Elimination Scoring (required)

Columns	Variable	Description
11-15	IELV	Criterion variable number
16-20	IEGP	Groupting parameter 1 = percent of subjects 2 = score range

Columns	Variable	Description		
21-30	REL(1)	LIMIT(1)	Low group	Middle gp
31-40	REL(2)	LIMIT(2)		
41-50	REL(3)	LIMIT(3)	High gp	
51-60	REL(4)	LIMIT(4)		
61-70	RETO	Tolerance for elimination scoring		

(6) Item Analysis Definition Card -- Classical Scoring (required)

Columns	Variable	Description		
11-15	ICLV	Criterion variable number		
16-20	ICGP	Grouping parameter 1 = percent of subjects 2 = score range		
21-30	RCL(1)	LIMIT(1)	Low group	Middle gp
31-40	RCL(2)	LIMIT(2)		
41-50	RCL(3)	LIMIT(3)	High group	
51-60	RCL(4)	LIMIT(4)		

(7) First Output Card (required)

Columns	Variable	Description
11-12	IØ(1)	Print input 0 = no 1 = long version 2 = short version
13-13	IØ(2)	Print and/or punch observed probabilities 0 = no 1 = print long version 2 = print short version 3 = print long version and punch all observed probabilities 4 = print short version and punch all observed probabilities 5 = punch all observed probabilities
15-16	IØ(3)	Print scores for each individual subject (0,1)

Columns	Variable	Description
17-18	IØ(4)	Print rosters of student scores 0 = no 1 = print minimum (10, INVAR) scores 2 = punch all INVAR scores 3 = both 1 and 2
19-20	IØ(9)	Item analysis indices for decision-theoretic scoring 0 = no 1 = using observed probabilities and observed log scores 2 = using adjusted probabilities and adjusted log scores 3 = both 1 and 2
21-22	IØ(10)	Item analysis indices for elimination scoring (0,1)
23-24	IØ(11)	Item analysis indices for classical scoring (0,1)
25-26	IØ(12)	Roster of item scores using observed probabilities 0 = no 1 = print 2 = punch 3 = both print and punch
27-28	IØ(13)	... using adjusted probabilities (0,1,2, Or 3)
29-30	IØ(14)	... using observed log scores (0,1,2, Or 3)
31-32	IØ(15)	... using adjusted log scores (0,1,2, or 3)
33-34	IØ(16)	... using elimination scores (0,1,2, or 3)
35-36	IØ(17)	... using classical scores (0,1,2, or 3)
37-38	IØ(18)	... using perceived information (0,1,2, or 3)
39-40	IØ(19)	... using actual information (0,1,2, or 3)

Columns	Variable	Description
41-42	IØ(20)	Reliability analysis using observed probabilities (0,1)
43-44	IØ(21)	... using adjusted probabilities (0,1)
45-46	IØ(22)	... using observed log scores (0,1)
47-48	IØ(23)	... using adjusted log scores (0,1)
49-50	IØ(24)	... using elimination scores (0,1)
51-52	IØ(25)	... using classical scores (0,1)
53-54	IØ(26)	... using perceived information (0,1)
55-56	IØ(27)	... using actual information (0,1)
57-58	IØ(28)	Summary of reliability analyses (including Livingston's coefficient) (0,1)

(8) Second Output Card (required)

Columns	Variable	Description
10-14	JØ(1)	These are subject variable numbers used for the rosters of subject scores; DEC-TEST expects INVAR of these variable numbers to be specified. If INVAR > 15, then use as many additional cards as may be required. Subsequent cards follow the same format as indicated here. The first variable on the first subsequent card would be JØ(16), the second JØ(17), etc.
15-18	JØ(2)	
19-22	JØ(3)	
23-26	JØ(4)	
27-30	JØ(5)	
31-34	JØ(6)	
35-38	JØ(7)	
39-42	JØ(8)	
43-46	JØ(9)	
47-50	JØ(10)	
51-54	JØ(11)	
55-58	JØ(12)	
59-62	JØ(13)	
63-66	JØ(14)	
67-70	JØ(15)	

(9) Third Output Card (required, but blank card may be used if IØ(28) = 0)

Columns	Variable	Description
1-5	CTT(1,1)	First criterion score for Livingston's Reliability Coefficient when reliability analysis uses observed probabilities
6-10	CTT(1,2)	Second ... observed probs.
11-15	CTT(2,1)	First ... adjusted probs.
16-20	CTT(2,2)	Second ... adjusted probs.
21-25	CTT(3,1)	First ... observed log scores
26-30	CTT(3,2)	Second ... observed log scores
31-35	CTT(4,1)	First ... adjusted log scores
36-40	CTT(4,2)	Second ... adjusted log scores
41-45	CTT(5,1)	First ... elimination scores
46-50	CTT(5,2)	Second ... elimination scores
51-55	CTT(6,1)	First ... classical scores
56-60	CTT(6,2)	Second ... classical scores
61-65	CTT(7,1)	First ... perceived information
66-70	CTT(7,2)	Second ... perceived info.
71-75	CTT(8,1)	First ... actual information
76-80	CTT(8,2)	Second ... actual information

(10) Object-Time Format Card(s) for Heading (required)

(11) Object-Time Format Card(s) for Subject Data Input (required -- use A and F format)

(12) Object-Time Format Card for Answer Key and Item Type Key (required, but blank card may be used if ITCDS = 1 -- use I format)

(13) Item Keys Definition Card (required)

Columns	Variable	Description
11-15	ITCDS	Item cards parameter 1 = one item card for each item 2 = answer key and item type key only
16-20	ITKEY	Order of answer key and item type key 1 = answer key first 2 = item type key first
21-25	ITSCO	Scores for each item (0,1)

Columns	Variable	Description
26-30	ITDEC	Decision-theoretic item analysis for each item 0 = no 1 = using observed probs. 2 = using adjusted probs. 3 = both 1 and 2
31-35	ITFLI	Elimination item analysis for each item (0,1)
36-40	ITCLA	Classical item analysis for each item (0,1)

Note: If ITCDS = 1, then the remaining paramoters are ignored by DEC-TEST and may be left blank.

(14, 15) Item Keys (required only if ITCDS = 2)

No keys should be present if ITCDS = 1.
Both keys must conform to format specified
by user in (12).
Both keys must be on logical unit LUCD
Answer Key comes first if ITKEY = 1.
Item Type Key comes first if ITKEY = 2.
Answer Key is IT(i,2), i = 1, 2, ..., K.
Item Type Key if is IT(i,3), i = 1, 2, ..., K

(16) Item Cards (required only if ITCDS = 1)

No item cards should be present if ITCDS = 2
Number of cards must equal K.
Cards must be on logical unit LUCD.
Each card must conform to the following format:

Columns	Variable	Description
7-10	JT(i)	User-defined item identification (A-format)
11-15	IT(i,2)	Correct answer (1 ≤ IT(i,2) ≤ IT(i,3))
16-20	IT(i,3)	Number of alternatives (2 ≤ IT(i,3) ≤ 5)
21-25	RIT(i)	Item weight -- if all items have equal weight, let RIT(i) = 1

Columns	Variable	Description
26-30	IT(i,4)	Split-halves key for reliability analyses 0 = skip item 1 = first half 2 = second half
31-35	IT(i,5)	Unweighted student scores on item (0,1)
36-40	IT(i,6)	Decision-theoretic item analysis 0 = no 1 = using observed probs. 2 = using adjusted probs. 3 = both 1 and 2
41-45	IT(i,7)	Elimination scoring item analysis (0,1)
46-50	IT(i,8)	Classical scoring item analysis (0,1)

(17) Student Data (required)

All data for one student constitute a student record which must conform to the format specified by the user in (11).
DEC-TEST expects N student records unless N = 0, in which case all student data is read until an end of file code, /*, is encountered.
All student data must be on logical unit LUSD.

(18) End of File Card (required)

The last card (or, preferably, the next to the last card) in the deck input to DEC-TEST should contain /* in columns 1 and 2, respectively, and blanks in the remaining 78 columns.

(19) End of Job Card (optional, but desirable)

Preferably, the last card in the deck input to DEC-TEST should contain // in columns 1 and 2, respectively, and blanks in the remaining 78 columns.

III. Detailed Description of Control Cards

The following pages provide a more detailed description of the parameters and variables defined by the user by means of the control cards.

(1) First Input Card (required)

RUN(5): a five-character alphanumeric run identification printed in the top left-hand corner of printed output and punched in the first five columns of punched output.

K: the number of items to be analyzed.

N: the number of student records input to DEC-TEST. N is not the number of cards containing student data. The Object-Time Format Card(s) for Subject Data Input defines one student record. If N = 0, then all student data is read until an end of file code is encountered, and DEC-TEST counts the number of student records.

INC: the number of columns for student identification. $0 \leq INC \leq 24$. If INC = 0, then the student data contains no student identification information, and DEC-TEST identifies students only according to a sequential student number; i.e., the first record of student data encountered is for student number 1, the second record for student number 2, ..., the last record encountered is for student number N.

INCS: the number of columns used for sorting student identifications. If student identifications are alphabetic (e.g., names), the sort results in an alphabetization for the number of columns specified. $0 \leq INCS \leq INC \leq 24$. if INCS = 0,then no sort is performed. If INC = 0, then INCS is ignored and may be left blank by the user.

IBGS: the first column for sorting the student identifications. Unless INC = 0 or INCS = 0, the columns sorted are columns IBGS to (IBGS + INCS - 1) in the student identifications. If INC = 0 or INCS = 0, then IBGS is ignored and may be left blank by the user. If INC $\neq$ 0 and INCS $\neq$ 0, then $1 \leq IBGS \leq INC \leq 24$ and $IBGS + INCS - 1 \leq INC$.

IXTRA: If IXTRA = 1, then the Object-Time Format Card(s) for Subject Data Input specifies an additional student variable input to DEC-TEST. This variable is available as a criterion variable for item analysis; it is treated just like any other student variable. If IXTRA = 0, then no additional student variable is included in the student input data.

XMS and MSD: XMS is the code used to identify missing data in the set of data input to DEC-TEST. XMS may be any real number consistent with the format specified by the Object-Time Format Card(s) for Subject Data Input; however, the user should be careful that XMS is not a valid score value for input. For example, if XMS is left blank, XMS is set to 0.0 by FORTRAN, but 0.0 may be a valid score value. Note that we are using the word "score" here to mean the score for alternative j on item i for student h. MSD is the technique for handling (processing) missing data for student h for item i. If MSD = 0 and the score values input for each of the n_i alternatives for item i equal XMS, then for the student under consideration, each alternative is assigned an observed probability of

$$p_{hij} = 1/n_i \ , \quad j = 1, \ , \ldots, n_i \ .$$

If MSD = 1 and the values input for each of the n_i alternatives for item i equal XMS, then, for the student under consideration, item i is skipped.

IOTH: the number of Object-Time Format Cards for Heading. $1 \leq \text{IOTH} \leq 2$.

IOTD: the number of Object-Time Format Cards for Subject Data Input. $1 \leq \text{IOTD} \leq 10$.

INVAR: the number of student variables on the Second Output Card(s). $0 \leq \text{INVAR} \leq 102$. If INVAR = 0, then any values on the Second Output Card are ignored; furthermore, stude variables numbered 6, 8, 62, 77. 72, 95, 90, 100, 99, and 101 are reported in the rosters of student scores if IØ(4) = 1 or 3 and/or punched out if IØ(4) = 2 or 3.

(2) Second Input Card (required)

IPT: the kind of student data used as input to DEC-TEST. The user can employ any one of five different kinds of input to DEC-TEST. The first three kinds of input are probabilities or linear transformations of probabilities, and the last two kinds are log scores. Note that DEC-TEST expects an input score (perhaps XMS) for each alternative, for each item, for each student. The kinds of input are as follows:

(a) If IPT = 1, then probabilities are used as input.

(b) If IPT = 2 , then probabilities multiplied by 100 are used as input.

(c) If IPT = 3, then the star method is used as input. In one of the original articles on decision-theoretic testing, de Finetti (1965) suggested that students could be told to respond to an item by distributing five stars or points over the set of alternatives for an item. If, for example, an item has five alternatives, and a student assigns one star or point to each alternative, then the probabilities associated with each alternative would be 0.20. DEC-TEST allows any total number of stars or points to be used for any item. That is, the total number of stars or points may differ for each item and/or for each student. DEC-TEST adds up the numbers (stars) assigned to each alternative for each item by each student and uses this total as the divisor for calculating probabilities for the item for the particular student.[1]

(d) If IPT = 4, then log scores are used as input. In general, the formula for a log score for student h on alternative j of item i is

$$L_{hij} = A_i \log R_{hij} + B_i$$

with a truncation value of C_i where

R_{hij} = original probability for student h, for item i, for alternative j.

[1] From a practical point of view, it is usually wise to tell the students that they should use a single, instructor-specified total number of points for all items on the test. Such a procedure simplifies the task for the students. Then, if any student uses a different total number of points for any item, by intent or by mistake, DEC-TEST will routinely make appropriate adjustments, as indicated above. The author has found that students readily understand this response strategy, whereas they sometimes have difficulty when asked to respond with log scores (IPT = 4 or 5) directly. Also, the author has found that, when the items on a test have two, three, or four alternatives, twelve points is a convenient number to use for distribution over the alternatives of an item. Since twelve is divisible by two, three, and four, students can always indicate "no knowledge" (i.e., a probability equal to $1/n_i$). Furthermore, twelve stars allow for a reasonably dense range of probabilities, and hence log scores.

Note that the original probabilities (designated by an upper case "R") are to be distinguished from the observed probabilities (designated by an upper case "P"); for further information about this distinction, see Section I and the discussion of DCT. Also, note the distinction between the log score for alternative j of item i, as given above, and the log score for item i, as given in Section I; i.e., the two log scores are equal only when j = * (the correct answer). The scoring function given above is actually a family of scoring functions, each member of which is determined once A_i, B_i, and C_i are specified (see Third Input Card). In DEC-TEST, these parameters must be the same for all items having the same number of alternatives.

(e) If IPT = 5, then the scores used as input are log scores in which $A_i = 50$, $B_i = 100$, and the truncation value $C_i = 0.01$ for all items, regardless of the number of alternatives. Furthermore, a score of 100 is input as 99. Using this procedure, each score for each alternative occupies no more than two positions (e.g., two card columns). Thus, this procedure is very useful when students use a SCoRule (a device for converting probabilities to log scores with $A_i = 50$, $B_i = 100$, and $C_i = 0.01$) to record log scores which are then punched on cards for input to DEC-TEST.

DCT: tolerance for decision theoretic testing. The first major computation performed by DEC-TEST involves converting (if necessary) the input student data to original probabilities R_{hij}. Note that, for this conversion procedure, any original probability less than the truncation value C_i is automatically converted to C_i. If we let Z_{hij} be a generic input score for student h, for item i, for alternative j, then:

(a) If IPT = 1, $R_{hij} = Z_{hij}$;

(b) If IPT = 2, $R_{hij} = Z_{hij}/100$;

(c) If IPT = 3, $R_{hij} = Z_{hij}/\sum_j Z_{hij}$;

(d) If IPT = 4, $R_{hij} = 10.0 \text{ EXP}[(B_i - Z_{hij})/A_i]$; and

(e) If IPT = 5, then whenever $Z_{hij} = 99$ <u>and</u> all other scores for item i, for student h are 0, change the "99" to "100" and use the formula for R_{hij} in (d), above, letting $A_i = 50$ and $B_i = 100$.

Now, theoretically,

$$\sum_{j=1}^{n_i} R_{hij} = 1.0$$

for each item, for each student. In practice, however, this is not always true; for example, students sometimes err in recording their responses, which results in the sum of the original probabilities for an item not equaling 1.0. If such errors are slight, then there is little cause for concern; however, large errors can be troublesome and usually indicate that a student is not following directions for recording responses. If, for student h on item i,

$$\text{ABS}(1.0 - \sum_{j=1}^{n_i} R_{hij}) > \text{DCT},$$

where ABS means "absolute value," then DEC-TEST calls this a validity check for student h and allows the user to perform a normalization procedure for the original probabilities. For example, suppose that DCT = 0.04 and the original probabilities are 0.00, 0.63, and 0.41. The probability 0.00 is automatically converted to 0.01; then, since

$$\text{ABS}\{1.0 - (0.01 + 0.63 + 0.41)\} = 0.05 > 0.04,$$

a validity check occurs.

NORM: the normalization procedure for items when a validity check occurs. Note that normalization is never performed for an item unless a validity check for that item occurs.

(a) If NORM = 0, then no normalization is performed and the original probabilities are simply called the observed probabilities.

(b) If NORM = 1, then normalization is performed over those alternatives for the item that are greater than the truncation value. For example, suppose C_i = 0.01, DCT = 0.04, NORM = 1, and the original probabilities for a three-alternative item are 0.01, 0.63, and 0.41. A validity check occurs (see discussion of DCT) and the original probabilities are transformed to 0.01, 0.63/1.04 = 0.6058, and 0.41/1.04 = 0.3942 . These new probabilities are called observed probabilities. Note that the sum of the observed probabilities not equal to 0.01 (the truncation value) is exactly 1.0 .

(c) If NORM = 2, then normalization is performed over all alternatives, with the constraint that none of the resulting observed probabilities is allowed to be less than the truncation value C_i. For example, if $C_i = 0.01$, DCT = 0.04, NORM = 2, and the original probabilities for a three-alternative item are 0.01, 0.63, and 0.41, then the observed probabilities become 0.01, $(0.63)(0.99)/1.04 = 0.5997$, and $(0.41)(0.99)/1.04 = 0.3903$. If, on the other hand, the original probabilitiet were 0.01, 0.55, and 0.42, then the observed probabilities would be $0.01/0.98 = 0.0102$, $0.55/0.98 = 0.5612$, and $0.4\,/0.98 = 0.4286$. Note that these new probabilities are called observed probabilities and they sum to exactly 1.0 .

LUCD: logical unit for reading item cards or item keys.

LUSD: logical unit for reading Student Data Input, including student identifications and the additional student variable, if present.

LUPT: primary logical unit for printing all output except that printed on LUPT2.

LUPT2: secondary logical unit for printing, which is used to print Individual Subject Scores and the Summary of Reliability Analyses.

LUPC: logical unit for punched output. Output designated as punched output can, of course, be written on any medium (e.g., cards, tape, disc, or drum); however, all "punched" records will be written as 80-column card images.

(3) Third Input Card (required)

The third input card reads in values for a matrix $AB(n_i,a)$, $n_i = 2,3,4$, or 5 (number of alternatives for an item) and a = 1,2, or 3 (parameters for log scoring function). Recall that, in general,

$$L_{hij} = A_i \log R_{hij} + B_i \ .$$

In terms of the matrix notation introduced above

$$B_i = AB(n_i,2),$$

$$C_i = AB(n_i,3), \text{ and}$$

$$A_i = -\,[B_i - AB(n_i,1)]/\log C_i \ .$$

From the last equation, it should be clear that

$$AB(n_i,1) = A_i \log C_i + B_i ;$$

i.e., the lowest score for the log scoring function for item type i (here, items having the same number of alternatives have the same item type) is obtained when R_{hij} is set to the truncation value C_i. Thus, once the user specifies the low score, high score, and truncation value, DEC-TEST has or can calculate the log scoring function parameters A_i, B_i, and C_i. For example, if

$$AB(n_i,1) = 0.0,$$

$$AB(n_i,2) = 100.0, \text{ and}$$

$$AB(n_i,3) = 0.01, \text{ then}$$

$$B_i = 100.0,\ C_i = 0.01, \text{ and}$$

$$A_i = -(100.0 - 0.0) \log 0.01 = 50.0 .$$

If the scoring system for item type i is such that $L_{hij} = 0.0$ when $R_{hij} = P_{hij} = 1/n_i$, then $AB(n_i,3)$ may be left blank and will be calculated by DEC-TEST using the formula:

$$AB(n_i,3) = 10.0 \text{ EXP} \left[\frac{AB(n_i,2) - AB(n_i,1)}{AB(n_i,2)}\right] \log\left(\frac{1}{n_i}\right)$$

Clearly, in this case, the lowest score should be negative. For example, if

$$AB(4,2) = 10.0 \text{ and } AB(4,1) = -10, \text{ then}$$

$$AB(4,3) = 10.0 \text{ EXP}[(20/10) \log (1/4)] = 0.0625,$$

$$A_4 = -20 / \log 0.0625 = 16.6096,$$

$$B_4 = 10, \text{ and } C_4 = 0.0625$$

If the test being analyzed does not make use of items having i alternatives, then $AB(n_i,1)$, $AB(n_i,2)$, and $AB(n_i,3)$ may be left blank.

(4-6) Item Analysis Definition Cards (required)

There are three kinds of Item Analysis Definition Cards. These cards define item analysis procedures when items are scored according to decision-theoretic scoring rules, elimination scoring rules, and classical scoring rules, respectively. The parameters defined by each card are identical in meaning, except for the inclusion of RETO (tolerance for elimination scoring) in the card for item analysis under elimination scoring rules.

Criterion variable. Any one of the 102 variables described in Section V may be chosen, by number, as the criterion variable for item analysis. The criterion variable functions in a manner similar to total test score in typical item analysis procedures. Variables numbered 77, 100, and 99 are common choices for decision-theoretic, elimination, and classical item analyses, respectively. Note that the "additional student variable" is available as a criterion variable for item analyses.

Grouping and limits. In order to perform item analyses, subjects must be assigned to groups in either one of two ways. If the grouping parameter equals 1, the-first step in the grouping procedure involves rank ordering students on the criterion variable. Then, the lowest

$$100.0[\text{LIMIT}(2) - \text{LIMIT}(1)]\%$$

of the students constitute the "lower" group, the next lowest

$$100.0[\text{LIMIT}(3) - \text{LIMIT}(2)]\%$$

of the students constitute the "middle" group, and the highest

$$100.0[\text{LIMIT}(4) - \text{LIMIT}(3)]\%$$

constitute the "upper" group. For example, if LIMIT(1) = 0.00, LIMIT(2) = 0.33, LIMIT(3) = 0.67, and LIMIT(4) = 1.00, then the lower group is the lowest 33% of the distribution, the middle group is the next lowest (middle) 34% of the distribution, and the upper group is the highest 33% of the distribution of students.

If the grouping parameter equals 2, then letting a generic score for student h be S_h, the lower group consists of all students for whom

$$\text{LIMIT}(1) <= S_h < \text{LIMIT}(2),$$

the middle group consists of all students for whom

$$\text{LIMIT}(2) <= S_h < \text{LIMIT}(3),$$

and the upper group consists of all students for whom

$$\text{LIMIT}(3) <= S_h <= \text{LIMIT}(4).$$

Note that when the groupinp parameter equals 2, the limits chosen must correspond to potential values of the criterion variable if the grouping procedure is to have meaningful results.

Whether the grouping parameter equals 1 or 2, any of the three groups can often be eliminated by setting that group's limits equal. Note, however, that if either the lower group or the upper group contains no students, then several of the item statistics usually reported cannot be calculated; in such cases a sequence of *'s is printed for the uncalculatable statistics.

RETO. RETO is the tolerance for elimination scoring. DEC-TEST must always have a value for RETO. If the user leaves RETO blank, then RETO is set to 0.0 automatically by FORTRAN. See Section I for a consideration of the function of RETO in elimination scoring. (Section I also describes the procedure for classical scoring given the observed probabilities.)

(7) First Output Card (required)

The First Output Card, as described in the Summary of Control Cards and Student Data -- Section II, is for the most part self-explanatory. Note, however, that "punch" should be interpreted as "write" on some medium (e.g., cards, tape, disc, or drum) defined by the user. For further, information see the opening paragraphs of Section IV and the descriptions of each of the various outputs.

(8) Second Output Card (required)

The Second Output Card lists, by number, the student variables to be printed and/or punched in the rosters of student scores (Output Nos. 6 and 36 in Section IV, respectively). The number of student variables expected by DEC-TEST is specified by INVAR on the First Input Card.

If INVAR = 0, then the Second Output Card may be left blank.

Note that 0 <= INVAR <=102, but only fifteen subject variables can fit on one card; thus, more than one card may br required. Each required additional card follows the same format as the original Second Output Card. For example, if INVAR = 20, then two cards are required -- one containing the first fifteen subject variable numbers and a subsequent card containing the last five subject variable numbers in the first five fields. Also, note that the order in which the subject variable numbers are specified is immaterial.

If INVAR = 0, then, as a default, the subject variables specified by DEC-TEST are numbers 6,8, 62, 77, 72, 95, 90, 100, 99, and 101. These variables will be punched out if IØ (4) = 2 or 3 and/or they will be printed (with verbal identifications as opposed to numerical identifications) if IØ(4) = 1 or 3.

If INVAR ≠ 0, then all INVAR variables will be punched out if IØ(4) = 2 or 3. If IØ(4) = 1 or 3 and if INVAR > 10, only the first 10 subject variables specified will be printed in the rosters of subject scores. If IØ(4) = 1 or 3 and 0 < INVAR <= 10, all subject variables specified will be printed.

(9) Third Output Card (required)

DEC-TEST allows the user to specify two criterion scores for calculating Livingston's Reliability Coefficient for each of the eight possible types of reliability analyses available in DEC-TEST.

See Output No. 34 ("Summary of Reliability Analyses") in Section IV for a discussion of Livingston's Reliability Coefficient. This is the only output that provides Livingston's Coefficients; these coefficients are not contained in Output Nos. 25-32 ("Reliability Analyses").

Livingston's Coefficients will be printed (and, therefore, two criterion scores are required) only if IØ(28) = 1 and the corresponding Reliability Analysis (controlled by IØ(20) to IØ(27)) is requested. If IØ(28) = 0 and/or if IØ(20) to IØ(27) are all equal to 0, then the Third Output Card may be left blank and will be ignored.

Note that if DEC-TEST expects a criterion score and the user leaves such a score blank, then the criterion score is set to "0", which is an acceptable value for calculating Livingston's Reliability Coefficient.

(10) Object-Time Format Card(s) for Heading (required)

DEC-TEST expects IOTH (see First Input Card) Object-Time Format Cards for a Heading that is placed at the top of each page of printed output. In order to have a heading printed, the FORTRAN rules for object-time format statements must be followed. At a minimum, this card should contain the characters (1H). This card (these cards) should define only one line of print.

(11) Object-Time Format Card(s) for Subject Data Input (required)

DEC-TEST expects IOTD (see First Input Card) Object-Time Format Card(s) for describing a record of Subject Data Input. All of the data for a particular subject should be contained in one record, and FORTRAN rules for object-time format statements should be followed.

The principles that should be followed in specifying this object-time format are:

(a) In general, the data in a student record should be ordered as follows: subject identification, responses for first item, responses for second item, ..., responses for Kth item, additional student variable.

(b) If INC = 0, then no subject identification is expected. If IXTRA = 0, then no additional student variable is expected. (See First Input Card.)

(c) Student identification information is specified in A-format. The number of A-format characters must equal INC.

(d) Item responses and the additional student variable (if present) are specified in F-format.

(e) Note that the format must take into account responses for all alternatives to all items. For example, if one has a 25-item test and each item has four alternatives, then (25)(4) = 100 responses are expected for each student; thus, the object-time format must specify 100 responses, not 25, and each student record must contain 100 responses.

(12-16) Cards Providing Item Information (subset required)

These final control cards provide DEC-TEST with information concerning the characteristics of each item and the kinds of analyses for each item desired by the user. This information can be conveyed in two ways: (a) a "long description" that provides this information for each item individually or (b) a "short description" that provides similar information for all items at once.

Long Description. When the user wants to specify information for each item individually, then the user should let ITCDS = 1 in the Item Keys Definition Card and leave the remaining parameters on this card blank (if specified, these parameters are ignored anyway); leave blank the Object-Time Format Card for Answer Key and Item Type Key (if a format is specified, the format is ignored anyway); not include an Item Type Key or an Answer Key -- no cards for these keys, not even blank ones; and provide one Item Card for each of the K items. Each item card allows the user to specify for each item:

(a) a four character alphanumeric identification of the item that is printed on item analysis output. (A sequential item number is always printed on such output, whether or not the user specifies an alphanumeric item identification.)

(b) the correct answer for the item expressed as an integer between "1" and the number of alternatives for the item; i.e., 1 <=IT(i,2) <= IT(i,3). The correct answer must be specified for each item.

(c) the number of alternatives for the item expressed as an integer between "2" and "5"; i.e., 2 <= IT(i,3) <= 5. The number of alternatives must be specified for each item.

(d) the item weight which may be any real number. If the item weight is set to "0" or left blank, then, in effect, the item under consideration is excluded in the calculation of student scores.

(e) the split-halves parameter, which is used for reliability analyses, only. If this parameter is "1" or "2", then the item is placed in the first or second "half" of the test, respectively; if this parameter is "0", then the item is eliminated from consideration in reliability analyses. Note that the number of items in the first and second "halves" of the test need not be equal.

(f) four parameters for controlling whether or not the user desires each of the five possible analyses (or outputs) provided by DEC-TEST for each item. Each of these analyses (or outputs) is described in Section IV.

Note that DEC-TEST expects item cards on logical unit LUCD (see Second Input Card).

Short Description. The short description is "short" in that the K item cards are not required; however, cards 12-15 are required.

The Object-Time Format Card for Answer Key and Item Type Key must be specified according to the FORTRAN rules for object-time formats. The format should be specified using I-format, not F-format.

ITCDS in the Item Keys Definition Card should be set to "2", and the remaining parameters in the card must be specified. ITKEY allows the user to specify which comes first in the user's control cards -- the Item Type Key (ITKEY = 2) or the Answer Key (ITKEY = 1). ITSCØ, ITDEC, ITELI, and ITCLA are analogous to IT(i,5), IT(i,6), IT(i,7), IT(i,8) in the Item Cards. For example, if ITSCØ = 1, then Output No. 8 (Unweighted Student Scores on Item -- see Section IV) is printed for each and every item; the same result would occur if the long description were used and IT(i,5) = 1 for all K items.

The Item Type Key and the Answer Key are placed in the control card deck in the order specified by ITKEY. They both must conform to the Object-Time Format Card for Answer Key and Item Type Key, and both keys must be on logical unit LUCD (see Second Input Card).

Note that the short description does not allow the user to define alphanumeric item identifications, item weights, or split-halves. When the short description is used, the only item identification is a sequential item number (i.e., the first item in a student record is labelled item number 1, the second item is labelled item number 2, ..., the last item is labelled item number K); all item weights are set to "1.0"; and an odd-even split-halves is automatically provided.

IV. Output

DEC-TEST provides 44 different kinds of output. In this section we provide a brief description of these kinds of output in the order in which they are received by the user. This section is intended to be used along with sample output.

Note that Output Nos. 1-32 are printed on logical unit LUPT, Output Nos. 33 and 34 are printed on logical unit LUPT2, and Output Nos. 35-44 are punched on logical unit LUPC. Although we use the words "print" and "punch", the user controls the medium on which output is written. DEC-TEST has been programmed so that outputs on LUPT and LUPT2 occupy no more than 132 characters in width,and outputs on LUPC are 80-column card images.

Each of the different kinds of punched output is identified by a header card(s), and output printed on LUPT is identified by a page number in the upper right-hand corner. For all output, a sequence of *'s replacing the value of some score or variable indicates that the score or variable cannot be calculated for the particular set of data being analyzed. For example, if a-variable has a standard deviation of zero, the the correlation between this variable and some other variable cannot be calculated.

Also, note that all standard deviations and variances reported are biased estimates.

In the following pages we provide an output number (used only for the purposes of this manual), an output title, and a description for each of the 44 different kinds of output generated by DEC-TEST.

(1) Title Page -- always printed.

(2) Control Cards -- always printed

This is an interpreted pseudo-listing of the control cards input to DEC-TEST. Note in particular that the Item Type Key and Answer Key (used when ITCDS = 2) are not printed in the manner in which they are submitted to DEC-TEST. However, the information provided by these keys is contained in the lines that begin "DATA FOR ITEM NO"; also, when ITCDS = 2, the other values on these lines of printout are obtained from the Item Keys Definition Card or assigned automatically by DEC-TEST.

(3) Student Data: Input

This output is printed in its entirety if IØ(1) = 1. If IØ(1) = 2, then only the first three and last three student records are printed.

The total number of responses (in the sense of alternatives) for each student is:

$$\sum_{i=1}^{K} n_i \ .$$

Thus, for example, if there are 25 items each with four alternatives, then the 100 responses made by each student will be printed on nine lines (12 responses on the first eight lines, and 4 responses on the last line). If IXTRA = 1, then the additional student variable will be the last value printed for each student.

The subject numbers reported are sequential subject numbers reflecting the order of the student records in the input data.

(4) Messages

This output is self explanatory. It is printed only if one or more students skip all items or all but one item.

(5) Student Data: Observed Probabilities

This output is printed in its entirety if IØ(2) = 1 or 3. If IØ(2) = 2 or 4, then only the first three and last three student records are printed.

If the Messages output has occurred, then certain subjects have been eliminated ard, therefore, do not appear here. In this case, the number of students for whom probabilities are reported will be less than N, which is, technically, the number of student records input to DEC-TEST. Since Output No. 5 constitutes the primary data matrix used to derive all future outputs, subsequent use of N in most outputs (exceptions are clear from the context of the output) refers to the reduced number of students. Note also that, if the Messages output occurs, the subject numbers indicated in Output No. 5 usually will not correspond exactly with those indicated in Output No. 3 (Student Data: Input).

The probabilities reported have been normalized, if a validity check occurred and normalization was requested. If MSD = 1, and if a student skipped an item, then XMS appears in place of an observed probability for each of the alternatives for the item for that student. If MSD = 0 and a student skipped item i, then $1/n_i$ appears as the observed probability for each of the alternatives for item i for that student.

The listing of the observed probabilities corresponds with the listing of Student Data Input (output No. 3), except for the fact that Student Data Input also contains the additional student variable if IXTRA = 1.

(6) Roster of Student Raw Scores -- printed if IØ(4) * 1 or 3

For information concerning the actual variables printed out see Section III -- Second Output Card(s).

If a sort was requested, this output will reflect the result of the sorting procedure. Subject numbers to the left of the parentheses are the "sorted subject numbers"; i.e., numbers that indicate students' positions in the sorted roster. Subject numbers within parentheses are the sequential subject numbers from Student Data Observed Probabilities (Output No. 5).

(7) Roster of Student z-Scores -- printed if IØ(4) = 1 or 3

This output corresponds with Output No. 6. The scores reported are non-normalized z-scores with a mean of 0.0 and a standard deviation of 1.0 .

(8) Unweighted Student Scores on Item i -- printed if ITSCO = 1 or IT(i,5) = 1

See Output No. 6 for a discussion of subject numbers. See Section I for a discussion of scores reported. Note that, if MSD = 1, then the sample size reported includes only those students who did not skip item i. This output provides the raw data for Output Nos. 9-12. The symbol "*" next to an alternative in Output Nos. 8-12 indicates the correct answer. Also, for these outputs, the first alternative is labelled "A", the second "B", etc.; the item numbers reported are sequential item numbers; and beside the sequential item numbers, in parentheses, are the user-defined item identifications, if any.

(9) Item Analysis for Item No. i Using Decision-Theoretic Scoring with Observed Probabilities -- printed if ITDEC = 1 or 3, or if IT(i,6) = 1 or 3

This output is divided into four parts and uses the data given in Output No. 8. The four parts are:

(a) Item Analysis Table. This is a three dimensional frequency distribution in which each entry represents the number of students in a particular group (lower, middle, upper, or total -- see Section III, Item Analysis Definition Cards) who responded with observed probabilities in a particular interval, for a particular alternative.

Let N'_1, N'_2, N'_3, and N'_4 be the number of subjects in the lower, middle, upper, and total groups, respectively. Now, the means and standard deviations below this table are based upon the observed probabilities not classified into intervals. Thus, for example, the mean observed probability for students in group g on alternative j is:

$$P_{.(g)j} = \frac{1}{N'_g} \sum_{h=1}^{N'_g} P_{h(g)j} \text{ , where}$$

$j = 1, 2, \ldots, n_i$ alternatives for item i,

$g = 1, 2, 3, 4$ groups, and

$h = 1, 2, \ldots, N'_g$ subjects in group g.

(b) Item Analysis Indices. The formulas for each of the nine indices are given below. Here (and elsewhere in this manual) j = * refers to the correct answer.

(1) Arithmetic mean item score using observed probabilities:

$AMP = P_{.(4)*}$, $0 <= AMP <= 1$.

(2) Difference discrimination index for arithmetic means using observed probabilities:

$DDAP = P_{.(3)*} - P_{.(1)*}$, $-1 <= DDAP <= 1$.

(3) Geometric mean item score using observed probabilities:

$$GMP = [\prod_{h=1}^{N_4'} P_{h(4)*}] \; EXP(1/N_4') \; , \; 0 <= GMP <= 1.$$

(4) Difference discrimination index for geometric means using observed probabilities:

$$DDGP = \{[\prod_{h=1}^{N_3'} P_{h(3)*}] \; EXP(1/N_3')\}$$

$$- \{[\prod_{h=1}^{N_1'} P_{h(1)*}] \; EXP(1/N_1')\} \quad ,$$

where -1 <= DDGP <= 1.

(5) Correlational discrimination index using observed probabilities (CDP): the Pearson product moment correlation coefficient between observed probabilities associated with the correct answer and scores on the criterion variable. (-1 <= CDP <= 1.)

(6) Average information (AVI): the arithmetic mean of the perceived information for each student reported in Output No. 8. (0 <= AVI <= $\log_2 n_i$.)

(7) Arithmetic mean item score using observed log scores:

$$AML = \frac{1}{N_4'} [\sum_{h=1}^{N_4'} (A_i \log P_{h(4)*} + B_i)]$$

$$= A_i \log GMP + B_i$$

$$= \frac{1}{N_4'} \sum_{h=1}^{N_4'} L_{h(4)*}$$

$$= L_{.(4)*} \quad ,$$

where, in general,

$L_{h(g)*}$ = log score for student h in group g for the correct answer to item i, and

$A_i \log C_i + B_i <= AML <= B_i$

(8) Differnece discrimination index for arithmetic means using observed log scores:

$$DDAL = [L_{.(3)*} - L_{.(1)*}]/(-A_i \log C_i) ,$$

where the denominator is the range of the log scoring function, and, therefore, -1 <= DDAL <= 1.

(9) Correlational discrimination index using observed log scores (CDL): the Pearson product moment correlation coefficient between observed log scores associated with the correct answer and scores on the criterion variable. (-1 <= CDL <= 1.)

(c) Pearson Product Moment Correlation Coefficients. These are correlations between the probabilities associated with all possible pairs of alternatives for each of the four groups.

(d) Frequency Distribution of Perceived Information. Note that the limits of the class intervals for information vary depending upon the number of alternatives that the item has.

(10) Item Analysis for Item No. i Using Decision-Theoretic Scoring with Adjusted Probabilities -- printed if ITDEC = 2 or 3, or if IT(i.) = 2 or 3

The format for this output is identical to that for Output No. 9. For an explanation of Output No. 10, merely make the following replacements in the explanation for Output No. 9: "observed" becomes "adjusted"; "perceived" becomes "actual"; "P" becomes "$\hat{P}$"; and "L" becomes "$\hat{L}$".

(11) Item Analysis for Item No. i Using Elimination Scoring -- printed if ITELI = 1 or IT(i,7) = 1

This output is divided into three parts and uses the data in Output No. 8. The three parts are:

(a) Item Analysis Table. The table is self-explanatory, with the exception of average item score, "AVE IT SC", for group g which is:

$$ASE_g = \sum_{h=1}^{N'_g} E_{h(g)} \text{ , where}$$

$E_{h(g)}$ = elimination score for student h in group g (for item i)

g = 1,2,3,4 groups (lower, middle, upper, and total, respectively), and

N'_g = number of students in group g.

(b) Item Analysis Indices. The formulas for each of the four indices are given below.

(1) Average item score:

$$ASE = ASE_4 \text{ , } -1 <= ASE <= 1.$$

(2) Standard deviation of item scores:

$$SDE = \sqrt{\frac{1}{N'_4-1} [\sum_{h=1}^{N'_4} E^2_{h(4)} - \frac{(ASE)^2}{N'_4}]} \text{ , } SDE >= 0.$$

(3) Difference discrimination index:

$$DDIE = (ASE_3 - ASE_1)/2.0, \quad -1 <= DDIE <= 1.$$

The denominator, 2.0, is the range of the possible elimination scores (unweighted) for an item.

(4) Correlational discrimination index (CDIE): the Pearson product moment correlation coefficient between elimination scores and criterion variable scores for item i. (-1 <= CDIE <= 1.)

(c) Frequency Distribution of All Possible Combinations of Eliminated Alternatives. This part of Output No. Il is self-explanatory.

(12) Item Analysis for Item No. i Using Classical Scoring -- printed if ITCLA = 1 or IT(i,8) = 1.

This output is divided into two parts and uses the data in Output No. 8. The two parts are:

(a) Item Analysis Table. This is, in essence, a standard item analysis table. The presence of fractional values can be explained through an example. Suppose that, for a three-alternative item, a student's observed probabilities are, in order, 0.40, 0.40, and 0.20. The student's classical score will be 0.50, 0.50, or 0.00 depending upon whether the first, second, or third alternative is the correct answer. Thus, 0.50 is added to the frequency counts for the first and second alternatives in order to provide our "best guess" concerning the number of students who would choose each alternative if students were forced to pick one and only one alternative.

(b) Item Analysis Indices. Let us express the average item score for students in group g as:

$$ASC_g = \frac{1}{N'_g} \sum_{h=1}^{N'_g} C_{h(g)} \quad \text{, where}$$

$C_{h(g)}$ = classical score for student h in group g,

g = 1, 2, 3, 4 groups (lower, middle, upper, and total, respectively), and

N'_g = number of students in group g.

The formulas for the four indices are as follows:

(1) Average item score:

$$ASC = ASC_4 \text{ ,} \quad 0 <= ASC <= 1.$$

(2) Standard deviation of item scores:

$$SDC = \sqrt{\frac{1}{N'_4-1} [\sum_{h=1}^{N'_4} C^2_{h(4)} - \frac{(ASC)^2}{N'_4}]} \text{ ,} \quad SDC >= 0.$$

(3) Difference discrimination index:

$DDIC = ASC_3 - ASC_1$, -1 <= DDIC <= 1.

(4) Correlational discrimination index (CDIC): the Pearson product moment correlation coefficient between classical scores and criterion scores for item i. (-1 <= CDIC <= 1.)

(13) Item Analysis Indices for Decision-Theoretic Scoring Using Observed Probabilities and Log Scores -- printed if IØ(9) = 1 or 3

Indices are printed for item i only if ITDEC = 1 or 3, or if IT(i,6) = 1 or 3. See Output No. 9 for a description of the indices reported.

(14) Item Analysis Indices for Decision-Theoretic Scoring Using Adjusted Probabilities and Log Scores -- printed if IØ(9) = 2 or 3

Indices are printed for item i only if ITDEC = 2 or 3, or if IT(i,6) = 2 or 3. See Output No. 10 for a description of the indices reported.

(15) Item Analysis Indices for Elimination Scoring -- printed if IØ(10) = 1

Indices are printed for item i only if ITELI = 1 or IT(i,7) = 1. See Output No. 11 for a description of the indices reported.

(16) Item Analysis Indices for Classical Scoring -- printed if IØ(11) = 1

Indices are printed for item i only if ITCLA = 1 or IT(i,8) = 1. See Output No. 12 for a description of the indices reported.

(17-24) Rosters of Students by Weighted Item Scores

Output No.	Data Used	Printed if
17	Observed Probabilities	IØ(12) = 1 or 3
18	Adjusted Probabilities	IØ(13) = 1 or 3
19	Observed Log Scores	IØ(14) = 1 or 3
20	Adjusted Log Scores	IØ(15) = 1 or 3
21	Elimination Scores	IØ(16) = 1 or 3
22	Classical Scores	IØ(17) = 1 or 3
23	Perceived Information	IØ(18) = 1 or 3
24	Actual Inofrmation	IØ(19) = 1 or 3

Note that these rosters report weighted item scores. Formulas for calculating the unweighted components of these scores are found in Section I. The unweighted scores are designated as P_{hi*}, $\hat{P}_{hi*}$, L_{hi*}, $\hat{L}_{hi*}$, E_{hi}, C_{hi}, I_{hi}, and $\hat{I}_{hi}$, respectively. Their weighted counterparts are w_iP_{hi*}, $w_i\hat{P}_{hi*}$, etc. Clearly, if $w_i = 1.0$ for i = 1, 2, ..., K, then unweighted scores and weighted scores are identical.

Ten scores are printed on each line. The first score reported is for item number 1, the second score for item number 2, etc. If MSD = 1 and an item was skipped by a student, then 999.999 is printed to indicate missing data. See Output No. 6 for a discussion of subject numbers.

(25-32) Reliability Analyses

Output No.	Data Used	Printed if
25	Observed Probabilities	IØ(20) = 1
26	Adjusted Probabilities	IØ(21) = 1
27	Observed Log Scores	IØ(22) = 1
28	Adjusted Log Scores	IØ(23) = 1
29	Elimination Scores	IØ(24) = 1
30	Classical Scores	IØ(25) = 1
31	Perceived Inofrmation	IØ(26) = 1
32	Actual Inofrmation	IØ(27) = 1

Each of the above outputs provides a Hoyt Analysis of Variance Reliability Analysis as well as a Split-Halves Analysis. In addition, for Output Nos. 25 and 26, DEC-TEST provides a Split Halves Analysis where the student score equals the geometric mean of the probabilities associated with correct answers.

If MSD = 1 and any item scores are missing (identified as 999.999 in Output Nos. 17-24), then, for Output Nos. 25-32, all missing item scores are transformed to the item scores that would result if $P_{hij} = 1/n_i$. In effect, this transformation has the same effect on Output Nos. 25-32 as setting MSD = 0 in the First Input Card.

In the following paragraphs, we explain the reliability analysis output, in general, and provide selected formulas. For these purposes let us define

X_{hi} = unweighted generic item score for student h on item i, where

$h = 1, 2, \ldots, N$ and

$i = 1, 2, \ldots, K'$.

One caveat is in order. If the split-halves parameter for item i, IT(i,4), equals "0", then, for reliability analyses, item i is skipped. In this case, the actual number of items used for reliability analyses will be less that K (the number of items input to DEC-TEST); hence, we use K' as the total number of items under consideration here.

Hoyt Analysis of Variance. Many discussions of Hoyt's (1941) procedure for calculating reliability are available. See, for example, Guilford (1954) which also provides an excellent treatment of most of the coefficients and scores reported in Output Nos. 25-32.

The general element of the matrix that forms the raw data for calculating Hoyt's Reliability Coefficient is $w_i X_{hi}$. The coefficient itself, which is identical to Cronbach's (1951) Coefficient Alpha, is:

$$r_{tt}(\text{Hoyt}) = 1 - \frac{\text{Mean Square (Remainder)}}{\text{Mean Square (Examinees)}} .$$

Now, to explain the way in which standard errors of measurement are reported in Reliability Analyses, let

$X_{h+} = \sum_{i=1}^{K'} w_i X_{hi}$ = weighted total score for student h,

$X_{h.} = X_{h+} / \sum_{i=1}^{K'} w_i$ = weighted mean score for student h,

$SD(X_{h+})$ = standard deviation of weighted student total scores, and

$SD(X_{h.})$ = standard deviation of weighted student mean scores.

In general, the standard error of measurement is:

$$SEM = s\sqrt{1 - r_{tt}} \quad .$$

When "s" is replaced by $SD(X_{h+})$ in the above equation, we get the standard error of measurement for student total scores, which is printed to the left of the parentheses in Output Nos. 25-32. When "s" is replaced by $SD(X_{h.})$ we get the standard error of measurement for student mean scores, which is printed within parentheses. All standard errors of measurement are reported by DEC-TEST in a similar manner.

Split-Halves. If IT(i,4) = 1, then item i goes in the first "half"; if IT(i,4) = 2, then item i goes in the second "half." In the table "STUDENT SCORE = SUM OVER ITEMS" is analogous to X_{h+} and "STUDENT SCORE = MEAN OVER ITEMS" is analogous to $X_{h.}$. Thus, $SD(X_{h+})$ and $SD(X_{h.})$ are referred to as the "STANDARD DEVIATION" for the "TOTAL" test when "STUDENT SCORE = SUM OVER ITEMS", and the "STANDARD DEVIATION" for the "TOTAL" test when "STUDENT SCORE = MEAN OVER ITEMS", respectively.

Now, let

r = correlation between two halves,

s_{d+} = standard deviation of differences using total student scores (placed to left of parentheses in output)

$s_{d.}$ = standard deviation of differences using student mean scores (placed within parentheses in output),

a = proportion of total item weights in first "half", and

$b = 1 - a$ = proportion of total item weights in second "half".

Using this notation, Horst's Reliability for Parts of Unequal Length can be expressed as:

$$r_{tt}(\text{Horst}) = \frac{r[\sqrt{r^2 + 4ab(1 - r^2)} - r]}{2ab(1 - r^2)} \quad .$$

Rulon's, Flanagan's, or Guttman's Reliability is:

$$r_{tt}(\text{Rulon}) = 1 - \{s_{d'}^2 \; / \; [SD(X_{h+})]^2\} \; .$$

Split-Halves where Student Score Equals Geometric Mean of Probabilities Associated with Correct Answer. Since, the subject score is a (geometric) mean, all entries in the output that depend upon "STUDENT SCORE = SUM OVER ITEMS" are filled with *'s. The geometric mean score for student h on the total test is:

$$\{ \prod_{i=1}^{K'} [X_{hi} \; EXP(w_i)]\} \; EXP(1 \; / \sum_{i=1}^{K'} w_i) \; ,$$

where X_{hi} is replaced by P_{hi*} or $\hat{P}_{hi*}$ depending upon whether one is considering Output No. 25 or 26, respectively. Similar formulas can be constructed to calculate a student's score for the first and second "half" tests. Rulon's Reliability Coefficient is meaningless for this kind of data, and, therefore, all results depending upon it are replaced by *'s. The user should be aware that the validity of using geometric means in a split-halves analysis is questionable, at best.

(33) Individual Subject Scores -- printed if IØ(3) = 1

DEC-TEST provides 102 scores for each individual subject, which are identified, in general, as VAR(1) to VAR(102). The most important of these scores are treated in Section I; all scores are considered in Section V and formulas for such scores are provided. Note that this output is printed on logical unit LUPT2, and page numbers are not provided. The calculations that produce this output are performed just prior to the printing of Output No. 6.

(34) Summary of Reliability Analyses (with addition of Livingston's Coefficients).

This output is printed if IØ(28) = 1 and at least one of the parameters IØ(20) to IØ(27) equal 1. Furthermore, a summary is provided only if the corresponding "complete" Reliability Analysis was requested.

The "USER DEFINED CUT-1" and USER DEFINED CUT-2" values are the criterion scores supplied by the user in the Third Output Card. These values are used to calculate Livingston's (1972) Reliability Coefficient defined as:

$$r_{tt}(\text{Liv}) = \frac{r_{tt}\ V(X) \ + \ (\bar{X} - C)^2}{V(X) \ + \ (\bar{X} - C)^2} \text{, where}$$

r_{tt} = any one of the reliability coefficients reported in Output Nos. 25-32,

C = criterion score for Livingston's Reliability Coefficient,

$\bar{X}$ = mean, over subjects, of scores, and

$V(X)$ = variance, over subjects, of scores.

Now, reliability is, in general, unaffected by whether the underlying student score is X_{h+} or $X_{h.}$. (See discussion of Reliability Analyses, Output Nos. 25-32 for the notation used here.) However, for this output we use $X_{h.}$ as the raw score for calculating $\bar{X}$ and for defining C ("CUT-1" or "CUT-2" in output); thus, we use:

$$\bar{X} = X_{..} = \frac{1}{N} \sum_{n=1}^{N} X_{h.} \text{ , and}$$

$$V(X) = [SD(X_{h.})]^2 \ .$$

This choice results in $\bar{X}$ having clearly defined limits. Specifically,

when the scores used are	the limits of $\bar{X}$ are
observed probabilities	$0 <= \bar{X} <= 1$
adjusted probabilities	$0 <= \bar{X} <= 1$

when the scores used are	the limits of $\bar{X}$ are
observed log scores[1] or adjusted log scores[1]	$A_i \log C_i + B_i <= \bar{X} <= B_i$
elimination scores	$-1 <= \bar{X} <= 1$
classical scores	$0 <= \bar{X} <= 1$
perceived information[2] or actual information[2]	$0 <= \bar{X} <= \dfrac{\sum_{i=1}^{K'} w_i \log_2 n_i}{\sum_{i=1}^{K'} w_i}$

[1] Technically, the limits provided are only an approximation for adjusted log scores for a test in which not all items are of the same item type (i.e., not all items have the same number of alternatives).

Furthermore, for both observed and adjusted log scores, the limits provided are exactly correct only if A_i, B_i, and C_i are identical for each item. If this is not true, then the lower limit is:

$$[\sum_{i=1}^{K'} w_i (A_i \log C_i + B_i)] / \sum_{i=1}^{K'} w_i$$

and the upper limit is:

$$\sum_{i=1}^{K'} w_i B_i / \sum_{i=1}^{K'} w_i \quad .$$

The experience of the author indicates that, eventhough the limits provided in the body of the text may not be exactly correct, for a given test, these limits are almost always a good enough approximation for practical use.

[2] Technically, the upper limit provided is only an approximation for actual information for a test in which not all items are of the same item type; nevertheless, even in this case, the author's experience indicates that the limit provided in the body of the text is a good enough approximation for practical use.

Now, referring to the formula for r_{tt}(Liv), note that

$$r_{tt}(\text{Liv}) = r_{tt} \text{ if } \overline{X} = C \text{ ;}$$

i.e., Livingston alleges that his coefficient gives the reliability that would result if C were the mean of the test. Thus, when choosing potential values of C for the Third Output Card, the user should choose values within the limits reported above for the various interpretations of X.

In short, this output reports

$$r_{tt} = r_{tt}(\text{Liv}) \quad \text{when } C = \overline{X},$$

$$r_{tt}(\text{Liv}) \text{ when } C = \text{"CUT-1", and}$$

$$r_{tt}(\text{Liv}) \text{ when } C = \text{"CUT-2"}$$

for each of the different reliability coefficients (r_{tt}) reported in Output Nos. 25-32.

(35) Observed Probabilities - Punched -- punched if IØ(2) = 3, 4, or 5

Note that, if IØ(2) >= 3, then all observed probabilities for all subjects are punched. There is no provision for punching out observed probabilities for the first three and the last three students, only. Thus, this output is analogous to Output No. 5 when IØ(2) = 1 or 3.

The format for card output is as follows:

Columns	Description
1-5	Run identification, RUN(5)
6	Blank
7-30	Student identification
31-33	Sequential card number for student = SCN
34-35	Blank
36-40	Observed probability: (1)(SCN)
41-45	Observed probability: (2)(SCN)
⋮	⋮
76-80	Observed probability: (9)(SCN)

Note that each card contains a maximum of nine observed probabilities (punched using format F5.3).

(36) Roster of Student Raw Scores - Punched -- punched if IØ(4) = 2 or 3

For information concerning variables punched out, see Section III -- Second Output Card(s). Output No. 36 is similar to Output No. 6.

The format for card output is as follows:

Columns	Description
1-5	Run identification, RUN(5)
6	Blank
7-30	Student identification
31-34	Sorted student number
35	(
36-39	Sequential student number
40	)
41-43	Sequential card number for student = SCN
45-53	Score number: (1)(SCN)
54-62	Score number: (2)(SCN)
63-71	Score number: (3)(SCN)
72-80	Score number: (4)(SCN)

Note that each card contains a maximum of four scores. The scores are punched in the order indicated by the sequence of variables in the Second Output Card(s). Scores are punched using format F9.3 .

(37-44) Rosters of Students by Weighted Item Scores - Punched

DEC-TEST provides eight such punched rosters.

Output No.	Data Used	Punched If
37	Observed Probabilities	IØ(12) = 2 or 3
38	Adjusted Probabilities	IØ(13) = 2 or 3
39	Observed Log Scores	IØ(14) = 2 or 3
40	Adjusted Log Scores	IØ(15) = 2 or 3
41	Elimination Scores	IØ(16) = 2 or 3
42	Classical Scores	IØ(17) = 2 or 3
43	Perceived Information	IØ(18) = 2 or 3
44	Actual Information	IØ(19) = 2 or 3

These outputs are analogous to Output Nos. 17-24, respectively.

The format for card output is as follows:

Columns	Description
1-5	Run identification, RUN(5)
6	Blank
7-30	Student identification
31-34	Sorted student number
35	(
36-39	Sequential student number
40	)
41-43	Sequential card number for student = SCN
44-52	Item number: (1)(SCN)
53-61	Item number: (2)(SCN)
62-70	Item number: (3)(SCN)
71-79	Item number: (4)(SCN)

Note that each card contains a maximum of four scores (punched using format F9.3) .

V. Individual Subject Scores

DEC-TEST provides 102 Individual Subject Scores (for each student), labelled VAR(1) to VAR(102) in Table A-4. Many of these scores have been introduced in Section I, and formulas for all scores are provided later in this section. Note that all scores except VAR(1), VAR(101), and VAR(102) take item weights into account. Table A-5 provides the Individual Subject Scores Output (Output No. 33) for a hypothetical student using the illustrative data introduced in Section I.

For the purposes of discussion, we will divide the Individual Subject Scores Output into five parts, and discuss each part separately.

VAR(2) to VAR(16): Variables Relating to Reference Lines

The Ideal and Realism Lines have been discussed in Section I. The extent to which a student is unrealistic is indicated by VAR(8) as well as by

$$\text{VAR}(5) - \text{VAR}(6) = 1.0 - \text{VAR}(6) .$$

However, note that:

$\text{VAR}(8) >= 0.0$, whereas

$1.0 - \text{VAR}(6) = 0.0$ if student is completely realistic,

$1.0 - \text{VAR}(6) > 0.0$ if student is over-confident, and

$1.0 - \text{VAR}(6) < 0.0$ if student is under-confident.

Therefore, VAR(8) is a measure of the magnitude of unrealistic student performance; whereas, 1.0 - VAR(6) is a measure of both the magnitude and direction of unrealistic student performance. For the illustrative data, VAR(8) = 11.7499 degrees and 1.0 - VAR(6) = 1.0 - 0.65563 = 0.34437. Thus, this hypothetical student is over-confident.

The Base Line is the estimated "realism" line if the student had always assigned a probability of 1.0 to a single alternative for each item and 0.0 to the other alternatives. In a sense, therefore, the Base Line is the "realism" line for classical scoring as opposed to decision-theoretic scoring.

TABLE A-4

Variable Numbers
Scores for an Individual Subject

	INTERCEPT	SLOPE
IDEAL LINE	VAR(2)	VAR(5)
REALISM LINE	VAR(3)	VAR(6)
BASE LINE	VAR(4)	VAR(7)

	DEC. DEG.	DEG.	MIN.
ANGLE BETWEEN IDEAL LINE AND REALISM LINE	VAR(8)	VAR(11)	VAR(14)
ANGLE BETWEEN IDEAL LINE AND BASE LINE	VAR(9)	VAR(12)	VAR(15)
ANGLE BETWEEN REALISM LINE AND BASE LINE	VAR(10)	VAR(13)	VAR(16)

PROBABILITY INTERVAL	NO. TIMES USED	NO. TIMES CORRECT	PROPORTION CORRECT
0.0<=P< 0.1	VAR(17)	VAR(27)	VAR(37)
0.1<=P< 0.2	VAR(18)	VAR(28)	VAR(38)
0.2<=P< 0.3	VAR(19)	VAR(29)	VAR(39)
0.3<=P< 0.4	VAR(20)	VAR(30)	VAR(40)
0.4<=P< 0.5	VAR(21)	VAR(31)	VAR(41)
0.5<=P< 0.6	VAR(22)	VAR(32)	VAR(42)
0.6<=P< 0.7	VAR(23)	VAR(33)	VAR(43)
0.7<=P< 0.8	VAR(24)	VAR(34)	VAR(44)
0.8<=P< 0.9	VAR(25)	VAR(35)	VAR(45)
0.9<=P<=1.0	VAR(26)	VAR(36)	VAR(46)

	IDEAL LN	REAL. LN	BASE LN
AVERAGE S.S. OF DEVIATIONS FROM	VAR(47)	VAR(48)	VAR(49)

	OVER ALL ITEMS		PER ITEM	
	ACTUAL	PERCEIVED	ACTUAL	PERCEIVED
ENTROPY (UNCERTAINTY)	VAR(50)	VAR(53)	VAR(56)	VAR(59)
INFORMATION	VAR(51)	VAR(54)	VAR(57)	VAP(60)
MAX. POSSIBLE INFO.	VAR(52)	VAR(55)	VAR(58)	VAR(61)

COEFFICIENT OF BIAS = VAR(62)

	LG SC OVER ITEMS	LG SC PER ITEM	AR MN PROB. SCORE	GM MN PROB. SCORE
POSSIBLE IMPROVEMENT FROM:				
BETTER USE OF INFO.	VAR(63)	VAR(72)	VAR(81)	VAR(90)
MORE INFORMATION	VAR(64)	VAR(73)	VAR(82)	Var(91)
SCORE RESULTING FROM:				
BETTER USE OF INFO.	VAR(65)	VAR(74)	VAR(83)	VAR(92)
MORE INFORMATION	VAR(66)	VAR(75)	VAR(84)	VAR(93)
TOTAL POSSIBLE IMPROVEMENT	VAR(67)	VAR(76)	VAR(85)	VAR(94)
OBSERVED SCORE	VAR(68)	VAR(77)	VAR(86)	VAR(95)
HIGHEST POSSIBLE SCORE	VAR(69)	VAR(78)	VAR(87)	VAR(96)
SCORE STUDENT EXPECTS	VAR(70)	VAR(79)	VAR(88)	VAR(97)
SCORE FOR NO KNOWLEDGE	VAR(71)	VAR(80)	VAR(89)	VAR(98)

CLASSICAL SCORE = VAR(99) ELIMINATION SCORE = VAR(100)

NUMBER OF VALIDITY CHECKS = VAR(101)
NUMBER OF ITEMS SCORED = VAR(102)
ADDITIONAL STUDENT VARIABLE = VAR(1)

TABLE A-5

Illustrative Data:
Scores for Individual Subject

```
                         INTERCEPT     SLOPE
          IDEAL LINE      0.0         1.00000
          REALISM LINE    0.13775     0.65563
          BASE LINE       0.23333     0.41667

                                                    DEC. DEG.   DEG.   MIN.
 ANGLE BETWEEN IDEAL LINE AND REALISM LINE           11.7490     11.    45.
 ANGLE BETWEEN IDEAL LINE AND BASE LINE              22.3801     22.    23.
 ANGLE BETWEEN REALISM LINE AND BASE LINE            10.6302     10.    38.

      PROBABILITY      NO. TIMES      NO. TIMES      PROPORTION
       INTERVAL          USED          CORRECT         CORRECT
      0.0<=P <0.1         4.0            1.0           0.2500
      0.1<=P <0.2         0.0            0.0           0.0
      0.2<=P <0.3         3.0            1.0           0.3333
      0.3<=P <0.4         4.0            0.0           0.0
      0.4<=P <0.5         6.0            3.0           0.5000
      0.5<=P <0.6         3.0            1.0           0.3333
      0.6<=P <0.7         1.0            1.0           1.0000
      0.7<=P <0.8         0.0            0.0           0.0
      0.8<=P <0.9         2.0            2.0           1.0000
      0.9<=P<=1.0         2.0            1.0           0.5000

                                              IDEAL LN    REAL. LN     BASE LN
  AVERAGE S.S. OF DEVIATIONS FROM              0.1552      0.1127      0.0899

                            OVER ALL ITEMS                  PER ITEM
                         ------------------          -------------------
                         ACTUAL    PERCEIVED         ACTUAL    PERCEIVED
 ENTROPY (UNCERTAINTY)  11.6678      9.5571          1.1668      0.9557
 INFORMATION             1.2570      3.3677          0.1257      0.3368
 MAX. POSSIBLE INFO.    12.9248     12.9248          1.2925      1.2925

                 COEFFICIENT OF BIAS =   16.330

                                 LG SC        LG SC        AR MN        GM MN
                                 OVER          PER         PROB.        PROB.
                                 ITEMS         ITEM        SCORE        SCORE
POSSIBLE IMPROVEMENT FROM:
   BETTER USE OF INFO.           14.790        1.479       -0.041       0.029
   MORE INFORMATION             178.178       17.818        0.521       0.560
SCORE RESULTING FROM:
   BETTER USE OF INFO.          821.322       82.132        0.479       0.440
   MORE INFORMATION             985.210       98.521        1.041       0.971
TOTAL POSSIBLE IMPROVEMENT      192.968       19.297        0.480       0.589
OBSERVED SCORE                  807.032       80.703        0.520       0.411
HIGHEST POSSIBLE SCORE         1000.000      100.000        1.000       1.000
SCORE STUDENT EXPECTS           856.151       85.615        0.583       0.516
SCORE FOR NO KNOWLEDGE          805.462       80.546        0.417       0.408

CLASSICAL SCORE =     6.500                 ELIMINATION SCORE =      2.000
                 NUMBER OF VALIDITY CHECKS =      0.
                 NUMBER OF ITEMS SCORED =     8.
                 ADDITIONAL STUDENT VARIABLE = *******
```

VAR(17) to VAR(49): Distribution of Observed Probabilities

VAR(17) to VAR(46) constitute the distribution of observed probabilities collapsed into ten class intervals of length 0.10 . These variables, therefore, provide the grouped data version of the kind of information presented in Table A-3 for the illustrative data.

VAR(47) to VAR(49) provide the average sums of squares of the grouped data points about each of the three reference lines. VAR(47) and VAR(49) are of questionable utility; however, VAR(48) provides an indication of the extent to which the least squares Realism Line is a good fit for the observed probability data points.

VAR(50) to VAR(62): Information Theoretic Measures of Student Performance

This section of the output contains three parts: (a) information measures over all items (i.e., for the test); (b) information measures per item (i.e., for the average over items); and (c) the Coefficient of Bias.

Part (a) can be graphically displayed as the Information Square in Figure A-2, which can be interpreted in terms of the Arabian proverb:

He who knows and knows that he knows,
He is wise, follow him.

He who knows and knows not that he knows,
He is asleep, awaken him.

He who knows not and knows not that he knows not,
He is a fool, shun him.

He who knows not and knows that he knows not,
He is a child, teach him.

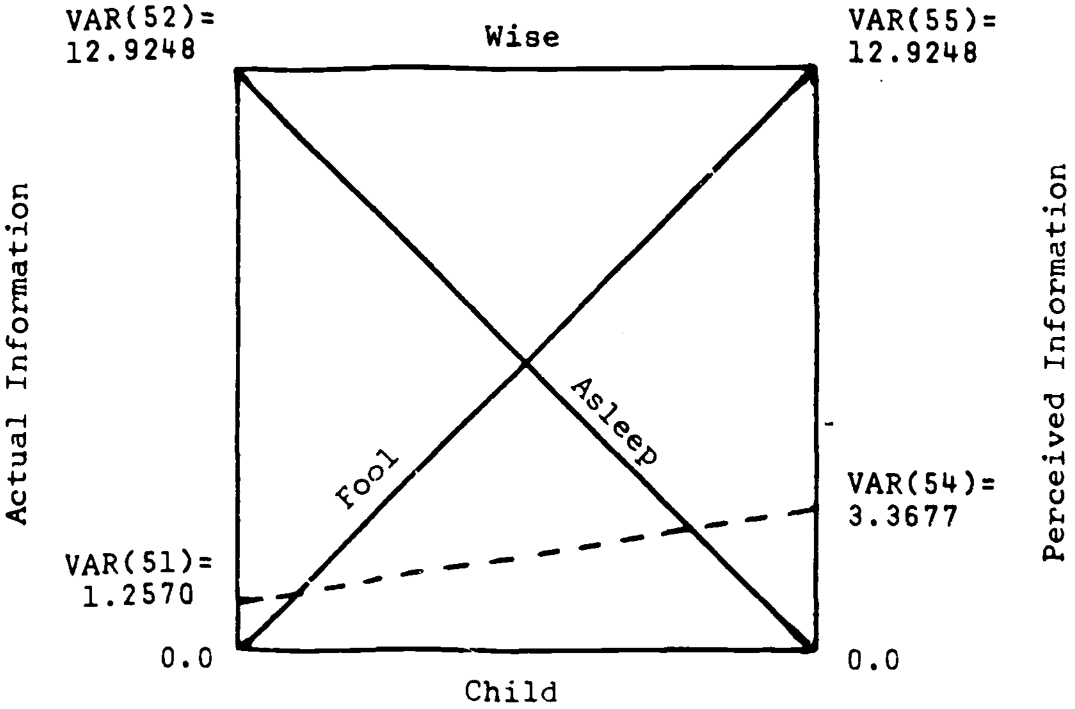
Since each variable in part (b) is a simple function of a corresponding variable in part (a), part (b) can also be graphically displayed in terms of the Information Square.

The Coefficient of Bias, VAR(62), provides another indication of the extent to which a student is unrealistic. Note that:

FIGURE A-2

Illustrative Data:

Information Square



Note.--The top horizontal line represents maximum possible actual and perceived information; the bottom horizontal line represents no information. The dashed line connects perceived and actual information. Numerical values reported are for the illustrative data. Be careful to grapg "information," not "entropy."

$-100.0 <= VAR(62) <= 100.0$

VAR(62) = 0.0 if student is completely realistic,

VAR(62) > 0.0 if student is over-confident, and

VAR(62) < 0.0 if student is under-confident.

For the illustrative data, VAR(62) = 16.330; therefore, the student is over-confident, which is consistent with the conclusion we reached when we observed that 1.0 - VAR(6) = 0.34437 > 0.0 . Another indication of over-confidence, on the part of our hypothetical student, is provided by the fact that the slope of the dashed line in the Information Square is greater than 0.0.

The user should be cautious in the interpretation of information and entropy measures in that the scale for these measures is non-linear (specifically, logarithmic -- base 2); hence, it is easy to fall into the error of over- and/or under-interpreting differences in magnitudes for these measures.[1]

VAR(63) to VAR(98): Primary Test Scores

This section of the output provides four sets of nine scores each, involving: (a) log scores over all items; (b) log scores per item (average over all items); (c) arithmetic mean probability scores; and (d) geometric mean probability scores. In general, parts (c) and (d) involve taking probabilities associated with the correct answers and calculating arithmetic and geometric weighted means,respectively. Note that all scores reported take item weights into account.

Parts (a), (b), and (d) provide essentially the same information using different measurement scales. Perhaps

[1] Another consideration is that, when not all items have the same number of alternatives, the maximum possible amount of actual information is not equal to the maximum possible amount of perceived information (see Section I). DEC-TEST handles this discrepancy by transforming actual information to the scale of perceived information (see formula for VAR(50)).

the most interesting scores reported are those that reflect a partitioning of total possible improvement (i.e., increase) in test score into improvement if the student makes better use of his or her information (i.e., if the student is more realistic) and improvement if the student had more information. Thus, in effect, the user can provide the student with quite detailed information concerning what the student might do to improve his or her score, as well as the potential effect of such action. Figure A-3 provides a student profile for selected geometric mean probability scores, for the illustrative data introduced in Section I. All nine geometric mean probability scores reported in the output are directly or indirectly represented in Figure A-3. Similar profiles could be constructed for the log scores in parts (a) and (b) of this section of the output.

Part(c) provides arithmetic mean probability scores. These scores are provided for comparative research purposes, only. Such scores are not appropriate for decision-theoretic testing, since they are based upon a linear scoring system. One glaring indication of this lack of appropriateness is that "POSSIBLE IMPROVEMENT FROM BETTER USE OF INFORMATION" is almost invariably a negative score indicating that students would almost always get lower (arithmetic mean probability) scores if they were more realistic.[1]

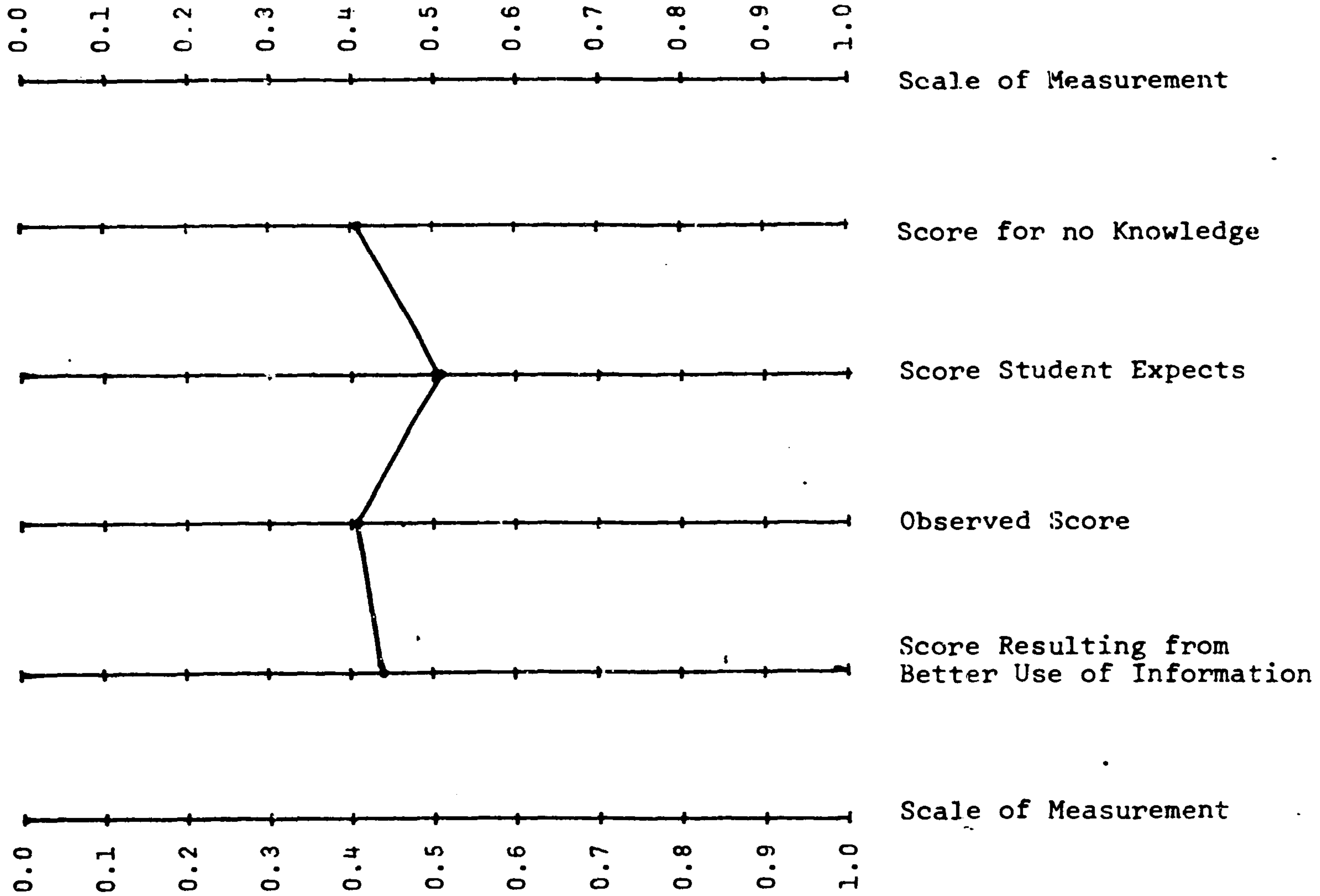
VAR(99) to VAR(102), VAR(1): Secondary Test Scores

The formulas for these scores are either self-explanatory or they provide a reference to an explanation.

[1] On rare occasions, "POSSIBLE IMPROVEMENT FROM BETTER USE OF INFORMATION" for parts (a), (b), and (c) may have a small negative value. Such negative values should be interpreted as 0.0 , since they are, for the most part, a result of rounding errors.

FIGURE A-3

Illustrative Data: Student Profile for Geometric Mean Probability Scores



Formulas for Individual Subject Scores

The following is a list of the formulas for the 102 Individual Subject Scores reported by DEC-TEST. There are, of course, a number of algebraically equivalent expressions for each of the equations listed here. For the most part, the actual equations provided are the ones actually used in programming DEC-TEST; however, these particular algebraic expressions may not always provide the most intuitively appealing definition of the variables. Thus, the user may wish to re-structure the algebraic expression for certain equations.

The formulas provided are listed in a sequential manner, according to the variable numbers; i.e., VAR(1), VAR(2), ..., VAR(102). However, the user should note that a particular VAR may be a function of a subsequently defined VAR; for example, VAR(64) is a function of VAR(67). This slight inconsistency is merely a result of the particular numbering scheme used to identify variables.

For the most part, the notational scheme used in the following formulas has already been introduced in Section I. The user should, however, note the following additions and minor modifications:

(a) the student subscript identifier, h, is dropped, since all formulas provide scores for one subject;

(b) the limits for subscripts i ($i = 1, 2, \ldots, K$) and j ($j = 1, 2, \ldots, n_i$) are not specified, since they remain constant;

(c) i' ($i' = 1, 2, \ldots, K$) is used as an additional item subscript;

(d) ℓ is used as a general purpose subscript, which is defined and/or given appropriate limits each time it is used;

(e) "INT" means "integer value";

(f) "ABS" means "absolute value";

(g) "EXP" means "exponential";

(h) "ATAN" means "arc-tangent expressed in radians"; and

(i) "*" is used as a multiplication operator as well as the indicator for correct answer.

VAR(1) = Score on Additional Student Variable

VAR(2) = 0.0

$$\text{VAR}(3) = \sum_i w_i\,[1.0 - \text{VAR}(6)] \,/\, \sum_i w_i n_i$$

$$\text{VAR}(4) = \sum_i w_i\,[1.0 - \text{VAR}(7)] \,/\, \sum_i w_i n_i$$

VAR(5) = 1.0

$$\text{VAR}(6) = \frac{\sum_i w_i P_{i*} - [(\sum_i w_i)^2 \,/\, \sum_i w_i n_i]}{\sum_i (w_i \sum_j P_{ij}^2) - [(\sum_i w_i)^2 \,/\, \sum_i w_i n_i]}$$

$$\text{VAR}(7) = \frac{\text{VAR}(99) - [(\sum_i w_i)^2 \,/\, \sum_i w_i n_i]}{\sum_i w_i - [(\sum_i w_i)^2 \,/\, \sum_i w_i n_i]}$$

$$\text{VAR}(8) = \text{ABS}\left\{\frac{180}{\pi}\ \text{ATAN}\left[\frac{1.0 - \text{VAR}(6)}{1.0 + \text{VAR}(6)}\right]\right\}$$

$$\text{VAR}(9) = \text{ABS}\left\{\frac{180}{\pi}\ \text{ATAN}\left[\frac{1.0 - \text{VAR}(7)}{1.0 + \text{VAR}(7)}\right]\right\}$$

$$\text{VAR}(10) = \text{ABS}\left\{\frac{180}{\pi}\ \text{ATAN}\left[\frac{\text{VAR}(6) - \text{VAR}(7)}{1.0 + \text{VAR}(6)*\text{VAR}(7)}\right]\right\}$$

VAR(11) = INT[VAR(8)]

VAR(12) = INT[VAR(9)]

VAR(13) = INT[VAR(10)]

VAR(14) = INT{[VAR(8) - VAR(11)]*60 + 0.5}

VAR(15) = INT{[VAR(9) - VAR(12)]*60 + 0.5}

VAR(16) = INT{[VAR(10) - VAR(13)]*60 + 0.5}

VAR(17) to VAR(26) = weighted number of times observed probability in given interval was used by student

VAR(27) to VAR(36) = weighted number of times probabilities in given interval were associated with correct answer

$$\mathrm{VAR}(36 + \ell) = \mathrm{VAR}(26 + \ell) \;/\; \mathrm{VAR}(16 + \ell) \;, \quad (\ell = 1, 2, \ldots, 10)$$

$$\mathrm{VAR}(47) = \sum_{\ell=1}^{10} \mathrm{VAR}(16+\ell)\{[\mathrm{VAR}(2)+\mathrm{VAR}(5)]*[(\ell/10)-0.05] - \mathrm{VAR}(36+\ell)\}^2 \;/\; \sum_i w_i n_i$$

$$\mathrm{VAR}(48) = \sum_{\ell=1}^{10} \mathrm{VAR}(16+\ell)\{[\mathrm{VAR}(3)+\mathrm{VAR}(6)]*[(\ell/10-0.05] - \mathrm{VAR}(36+\ell)\}^2 \;/\; \sum_i w_i n_i$$

$$\mathrm{VAR}(49) = \sum_{\ell=1}^{10} \mathrm{VAR}(16+\ell)\{[\mathrm{VAR}(4)+\mathrm{VAR}(7)]*[(\ell/10-0.05] - \mathrm{VAR}(36+\ell)\}^2 \;/\; \sum_i w_i n_i$$

$$\mathrm{VAR}(50) = \frac{[\sum_i(w_i \sum_j P_{ij}\log_2 P_{ij})]*[\sum_i w_i \log_2(n_i)]}{\sum_i w_i\{(n_i\alpha_h+\beta_h)[\log_2(n_i\alpha_h+\beta_h) - \log_2(n_i)]\}}$$

where α_h = VAR(3)

and β_h = VAR(6)

$$\mathrm{VAR}(51) = \mathrm{VAR}(52) - \mathrm{VAR}(50)$$

$$\mathrm{VAR}(52) = \sum_i w_i \log_2(n_i)$$

$$\mathrm{VAR}(53) = -\sum_i [w_i \sum_j P_{ij} \log_2(P_{ij})]$$

$$\mathrm{VAR}(54) = \mathrm{VAR}(55) - \mathrm{VAR}(53)$$

$$\mathrm{VAR}(55) = \mathrm{VAR}(52)$$

$$\mathrm{VAR}(55 + \ell) = \mathrm{VAR}(49 + \ell) / \sum_i w_i , \qquad (\ell = 1, 2, \ldots, 6)$$

$$\mathrm{VAR}(62) = \{[\mathrm{VAR}(54) - \mathrm{VAR}(51)] / \mathrm{VAR}(55)\}*100.0$$

$$\mathrm{VAR}(63) = \mathrm{VAR}(65) - \mathrm{VAR}(68)$$

$$\mathrm{VAR}(64) = \mathrm{VAR}(67) - \mathrm{VAR}(63)$$

$$\mathrm{VAR}(65) = \sum_i \{w_i \, [A_i \log(\hat{P}_{i*}) + B_i]\}$$

$$\mathrm{VAR}(66) = \mathrm{VAR}(68) - \mathrm{VAR}(64)$$

$$\mathrm{VAR}(67) = \mathrm{VAR}(69) - \mathrm{VAR}(68)$$

$$\mathrm{VAR}(68) = \sum_i \{w_i \, [A_i \log(P_{i*}) + B_i]\}$$

$$\mathrm{VAR}(69) = \sum_i w_i B_i$$

$$\mathrm{VAR}(70) = \sum_i w_i \, \{\sum_j P_{ij} \, [A_i \log(P_{ij}) + B_i]\}$$

$$\mathrm{VAR}(71) = \sum_i w_i \, [A_i \log(1.0/n_i) + B_i]$$

$$\mathrm{VAR}(71 + \ell) = \mathrm{VAR}(62 + \ell) / \sum_i w_i , \qquad (\ell = 1, 2, \ldots, 9)$$

VAR(96) = 1.0

VAR(97) = $10.0 \text{ EXP}\{[\sum_{i} w_i \sum_{j} P_{ij} \log(P_{ij})] / \sum_{i'} w_{i'}\}$

$= \prod_{i} \prod_{j} \{P_{ij} \text{ EXP } [w_i P_{ij} / \sum_{i'} w_{i'}\}$

VAR(98) = $\prod_{i} [n_i \text{ EXP } (-w_i / \sum_{i'} w_{i'})]$

VAR(99) = estimated weighted number of items correct if student forced to respond to each item with one and only one choice of correct answer. (See Section I.)

VAR(100) = estimated weighted elimination score for test. (See Section I.)

VAR(101) = unweighted number of validity checks. (See Section III -- DCT)

VAR(102) = unweighted number of items scored.

If MSD = 0 , VAR(102) = K ;
If MSD = 1 , VAR(102) <= K ;
i.e., an item is not scored if the input responses for all alternatives equal "XMS", the code for missing data.

VI. Technical Data and Information

Structure of DEC-TEST

DEC-TEST consists of a MAINLINE program, 16 subroutines, and a BLØCK DATA subprogram. Table A.6 lists selected technical characteristics of each program unit. The principal function of MAINLINE is the assignment.of user-defined values for modifiable assignment statements and modifiable dimension statements. INPUT serves as the principal program unit (subroutine) for reading control cards and student data, as well as for branching to other subroutines.

As indicated in Table A-6, DEC-TEST requires 124,294 bytes of main storage if no overlay structure is used. If the overlay structure indicated in Table A-6 and Figure A-7 is used, then DEC-TEST requires 70094 bytes of main storage; i.e., DEC-TEST requires the number of bytes necessary to store Segment 1 and Segment 5. Thus, the use of the overlay structure saves 124,294 - 70,094 = 54,200 bytes. However, these figures do not include:- (a) bytes required for user-defined matrices and vectors and (b) additional bytes (overhead) required by FORTRAN for execution of DEC-TEST.

User-Modifications of DEC-TEST

Figure A-4 provides a partial listing of the MAINLINE program for DEC-TEST. Both the modifiable dimension statements (MAI 7 to MAI 19) and the modifiable assignment statements (MAI 30 to MAI 33) can be altered by the user prior to compilation of DEC-TEST. Figure A-5 provides a worksheet for making such changes and determining the total number of bytes required by DEC-TEST. Note that, in order to execute DEC-TEST, the user needs additional bytes (overhead) required by FORTRAN; for this purpose, in most cases, 10,000 bytes should be more than sufficient.

In Figures A-4 and A-5

NDIM = maximum number of students for a run,

KDIM = maximum number of items for a run, and

IADIM = maximum number of responses (alternatives) for a student.

TABLE A-6

Structure of DEC-TEST

Program Unit[a]	Abbreviation[b]	Segment[c]	No. of Bytes[d]	No. of Source Statements[e]	No. of Cards[e]
MAINLINE	MAI	1	760	32	53
INPUT	INP	1	22516	382	524
PAGER	PAG	1	424	11	13
STDV	STD	1	472	8	10
CØRR	CØR	1	546	7	10
SEM	SEM	1	470	8	10
BLØCK DATA	BLK	1	25994	6	10
other		1		0	0
CØVER	CØV	2	2500	70	83
IØSDI	IØS	3	2768	58	65
SETUP	SET	4	4938	150	165
SCØRE	SCØ	5	18912	460	493
UML	UML	6	2352	63	75
IADCT	IAD	7	10292	258	295
IAELIM	IAE	8	6100	163	177
IACLAS	IAC	9	4400	111	123
SUMRY	SUM	10	3528	94	108
SXITEM	SXI	11	8994	257	277
RELIAB	REL	12	8328	244	268
		Totals:	124,294	2384	2769

[a] All "program units" are subroutines except for MAINLINE, BLØCK DATA, and "other." "Other" includes FORTRAN supplied subroutines, functions, , etc. required by DEC-TEST.

[b] These abbreviations are found in columns 72-74 of the source deck for DEC-TEST. Each card in the source deck is uniquely identified by the appropriate abbreviation followed by a sequential (within subroutine) card number in columns 76-80.

[c] The segment numbers refer to the overlay structure for DEC-TEST. During program execution, if the user employs the overlay structure, then main storage contains Segment 1 (root segment) and one of the Segments 2-12.

[d] The number of bytes required by user-defined matrices and vectors is not included here. The number of bytes for "other" includes FORTRAN supplied subroutines, functions, etc. required by DEC-TEST.

[e] "No. of cards" equals "no. of source statements" plus number of comment cards plus number of continuation cards.

FIGURE A-4

Partial Listing of MAINLINE Program for DEC-TEST

```
C MAINLINE PROGRAM FOR DECTEST:  A FORTRAN IV PROGRAM FOR DECISION-
C THEORETIC TEST SCORING AND THE ANALYSIS OF ITEM DATA WRITTEN BY
C ROBERT L. BRENNAN, DEPARTMENT OF EDUCATION, SUNY AT STONY BROOK
C
C MODIFIABLE DIMENSION STATEMENTS
C
      DIMENSION X(59,200)
      DIMENSION Z(51,20)
      DIMENSION W(59,20)
      DIMENSION RIT(51)
      DIMENSION T(50,32)
      DIMENSION RS(50)
      DIMENSION RT(50)
      DIMENSION JT(50)
      INTEGER*2 Y(59,24)
      INTEGER*2 IYSRT(59)
      INTEGER*2 IT(51,9)
      INTEGER*2 MS(50)
      INTEGER*2 IDDM(59)
C
C NON-MODIFIABLE COMMON STATEMENTS
C
      COMMON /COM1/ LUCC,IPAGE,NDIM,KDIM,IYC,IZC,IITC,ITALT
      COMMON /COM2/ IADIM,RN,RK,KDIMP1,IWC,NDIMP2,INVAR,MSC
      COMMON /COM4/ PERI(3),ACTI(3),IDUMMY(5),IRC,ITC
      COMMON /COM16/ ID(28),JD(101),AR(5,3),CORFAC,CTT(8,2)
C
C MODIFIABLE ASSIGNMENT STATEMENTS
C
      NDIM = 59
      KDIM = 50
      IADIM = 200
      LUCC = 5
```

FIGURE A-5

Worksheet for User-Defined Matrices, Vectors, and Assignment Statements

NDIM = ____ KDIM = ____ IADIM = ____ LUCC = ____

No. of bytes for user-defined matrices and vectors
= NDIM[(4)(IADIM) + 212] + (164)(KDIM) + 182
= ____[(4)(____) + 212] + (164)(____) + 182
= ______ bytes[a]

Variable Dimensions	User Dimensions	No. of Locations		No. of Bytes
X(NDIM,IADIM)	X(___,___)	______		
Z(NDIM+2,20)	Z(___, 20)	______		
R(NDIM,20)	R(___, 20)	______		
RIT(KDIM+1)	RIT(___)	______		
T(KDIM,32)	T(___,32)	______		
RS(KDIM)	RS(___)	______		
RT(KDIM)	RT(___)	______		
JT(KDIM)	JT(___)	______		
	Subtotal-1	______	X 4 =	______ bytes
Y(NDIM,24)	Y(___, 24)	______		
IYSRT(NDIM)	IYSRT(___)	______		
IT(KDIM+1,9)	IT(___, 9)	______		
MS(KDIM)	MS(___)	______		
IDDM(NDIM)	IDDM(___)	______		
	Subtotal-2	______	X 2 =	______ bytes

No. of bytes for user-defined matrices and vectors (Subtotal-1 + Subtotal-2) = ______ bytes[a]

Number of bytes required by DEC-TEST using overlay (70094 bytes) or not using overlay (124294 bytes) = ______ bytes

Total = ______ bytes

[a] These two values should be identical.

These are the only three variables required to define matrix and vector dimensions. Note that each element of the first eight matrices (or vectors) is a real variable occupying four bytes of main storage; while each element of the last five matrices (or vectors) is an integer variable occupying two bytes of main storage. Thus, the first eight matrices (or vectors) are associated with DIMENSIØN statements; while the last five matrices (or vectors) are associated with INTEGER*2 statements. Also, note that the number of elements (or locations) for a matrix is the product of its dimensions.

As indicated at the beginning of Section II, LUCC, the logical unit for reading (most of) the control cards, may be altered in the MAINLINE program.

Figure A-6 provides an example of the worksheet in Figure A-5. For this example, the total number of bytes required to execute DEC-TEST, using the overlay, is about 137984 + 10000 = 147984.

JCL for DEC-TEST at SUSB

Figure A-7 provides a listing of the JCL (Job Control Language) statements necessary to compile, link-edit, and execute DEC-TEST at the SUSB (State University of New York at Stony Brook) Computing Center. At SUSB logical unit numbers 5, 6, and 7 are defined in the catalogued procedure for FØRTGCLG as logical units for reading punched cards, printing, and punching, respectively. Any other required logical unit must be defined by the user with a //GØ.FT ... statement (see Fortran Programmer's Guide or JCL Manual).

If (in addition to compiling, linkediting, and executing DEC-TEST) one wanted to store a DEC-TEST load module (say D5950) in a catelogued dataset (say TESTAID) on a disk (say USER01), then the following statement would be placed immediately before the //LKED.SYSIN DD * card in Figure A-7:

```
//LKED.SYSLMØD DD DSN=USER.TESTAID(D5950),DISP=(ØLD,KEEP),
//     SPACE=(TRK,(5,5,2),RLSE),VØL=SLR=USER01,UNIT=3330
```

FIGURE A-6

Example of
Worksheet for User-Defined Matrices,
Vectors, and Assignment Statements

NDIM = 59   KDIM = 50   IADIM = 200   LUCC = 5

No. of bytes for user-defined matrices and vectors

= NDIM[(4)(IADIM) + 212] + (164)(KDIM) + 182

= 59[(4)(200) + 212] + (164)(50) + 182

= 68090 bytes[a]

Variable Dimensions	User Dimensions	No. of Locations	No. of Bytes
X(NDIM,IADIM)	X(59,200)	11800	
Z(NDIM+2,20)	Z(61, 20)	1220	
R(NDIM,20)	R(59, 20)	1180	
RIT(KDIM+1)	RIT(51)	51	
T(KDIM,32)	T(50,32)	1600	
RS(KDIM)	RS(50)	50	
RT(KDIM)	RT(50)	50	
JT(KDIM)	JT(50)	50	
	Subtotal-1	16001 X 4 =	64004 bytes
Y(NDIM,24)	Y(59, 24)	1416	
IYSRT(NDIM)	IYSRT(59)	59	
IT(KDIM+1,9)	IT(51, 9)	459	
MS(KDIM)	MS(50)	50	
IDDM(NDIM)	IDDM(59)	59	
	Subtotal-2	2043 X 2 =	4086 bytes

No. of bytes for user-defined matrices and vectors (Subtotal-1 + Subtotal-2) = 68090 bytes[a]

Number of bytes required by DEC-TEST using overlay (70094 bytes) or not using overlay (124294 bytes) = 70094 bytes

Total = 138184 bytes

[a] These two values should be identical.

FIGURE A-7

JCL to Compile, Linkedit, and Execute DEC-TEST with Overlay

```
              1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
1 3 5 7 9 1 3 5 7 9 1 3 5 7 9 1 3 5 7 9   Card Columns

// (job card)
// (account card)
//   EXEC FØRTGCLG,PARM.LKED='ØVLY'
//FØRT.SYSIN DD *

   (source deck)

/*
//LKED.SYSIN DD *
 ENTRY MAIN
 INSERT MAIN,INPUT,PAGER,STDV,CØRR,SEM
 ØVERLAY ALPHA
 INSERT CØVER
 ØVERLAY ALPHA
 INSERT IØSDI
 ØVERLAY ALPHA
 INSERT SETUP
 ØVERLAY ALPHA
 INSERT SCØRE
 ØVERLAY ALPHA
 INSERT UML
 ØVERLAY ALPHA
 INSERT IADCT
 ØVERLAY ALPHA
 INSERT IAELIM
 ØVERLAY ALPHA
 INSERT IACLAS
 ØVERLAY ALPHA
 INSERT SUMRY
 ØVERLAY ALPHA
 INSERT SXITEM
 ØVERLAY ALPHA
 INSERT RELIAB
/*
//GØ.FT__F001 DD SYSØUT=···
  :
  :
//GØ.FT__F001 DD SYSØUT=···
//GØ.FT05F001 DD *

   (DEC-TEST control cards and student data)

/*
//
```

Overlay--These statements in conjunction with PARM.LKED='ØVLY' on EXEC card accomplish the overlay.

FT cards--Positions underlined should be filled in with logical unit numbers required for user's run of DEC-TEST

Once this is done, the user can execute DEC-TEST using the following JCL statements:

```
// (job card)
// (account card)
//   EXEC PGM=D5950
//STEPLIB DD DSN=USER.TESTAID,DISP=ØLD
//FT__F001 DD SYSØUT=···
   .
   .
   .
//FT__F001 DD SYSØUT=···
//FT05F001 DD *

   (DEC-TEST control cards and student data)

/*
//
```

Bibliography

Coombs, C. H., Milholland. J. E. & Womer, F. B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 13-37.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334

de Finetti, B. Methods of discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 1965, 13, 87-123.

Echternacht, G. T. The use of confidence testing in objective tests. Review of Educational Research, 1972, 42, 217-236.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Hoyt, C. J. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153=160.

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

Savage, L. J. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association, 1971, 66, 783-801.

Shannon, C. E. & Weaver, W. The mathematical theory of communication. Urbana: The University of Illi Press, 1949.

Shuford, E. H., Albert, A. & Massengill, H. E. Admissible probability measurement procedures. Psychometrika, 1966, 31, 125-145.

APPENDIX B

Test Items Used for This Study

The following is a list of test items used in this study. An "*" beside an alternative indicates the correct answer. The following identification scheme for items is employed:

(a) Items identified as AC01 to AC25 are criterion-referenced items for test A, which was used as a pretest for some subjects, as a posttest for other subjects, and as both a pre- and posttest for still other subjects.

(b) Items identified as BC01 to BC25 are criterion-referenced items for test B, which was used a a pretest for some subjects, as a posttest for other subjects, and as both a pre- and posttest for still other subjects. Note that ACyy is intended to be equivalent to BCyy, where "yy" is any item number (01 to 25).

(c) Items identified as ZC26 to ZC50 are criterion-referenced items which all subjects took in the posttest mode, only. None of the ZC items are intended to be equivalent to any AC or BC item.

AC Items

AC01 Objectives have not been defined for which of the following domains?
A. affective
B. cognitive
* C. objective
D. psychomotor

AC02 Which of the following terms is least acceptable for instructional objectives which are to be measured through multiple-choice test items?
A. recognize
B. differentiate
C. identify
* D. list

AC03 Which of the following is most correct? Instructional objectives should:
A. be stated in terms of teacher behavior
B. end with an active verb
C. relate to one or two processes only
* D. represent intended direct outcomes of learning experiences

AC04 Which of the following is not correct? "Standardized" tests provide a standard for:
* A. excellence
B. timing
C. scoring
D. administration

AC05 The word "criterion" in "criterion-referenced test" usually refers to:
A. a cut-off value, such as 85% correct
* B. a set of objectives
C. some type of norms
D. another test

AC06 The assignment of numerals to objects or events according to rules is a difinition of:
* A. measurement
B. evaluation
C. Testing
D. Validation

AC07 In educational measurement, the underlying scale of measurement is usually:
A. nominal
* B. ordinal
C. interval
D. ratio

AC08 The percentage of students who get an item correct is called:
* A. difficulty level
B. error rate
C. theoretical difficulty level
D. theoretical error rate

AC09 A statistic used to show how sharply an item differentiates between the students who scored highest on a test and the students who scored lowest is called a (an):
A. difficulty level
B. error rate
* C. discrimination index
D. out-off value

AC10 Which of the following is not a possible value of the standard deviation of a set of scores?
A. 0.00
B. 100.03
C. 0.01
* D. -1.00

AC11 The 50th percentile is also called the:
A. standard deviation
B. mean
C. semi-interquartile range
* D. median

AC12 Which of the following is not a possible value for the Pearson product-moment correlation coefficient?
A. 0.00
B. 0.50
C. -1.00
* D. 1.25

AC13 If test scores are distributed normally, what percent of the scores will exceed a score falling one standard deviation below the mean?
A. 16%
B. 34%
C. 68%
* D. 84%

AC14 Which of the following estimates of reliability is most clearly assiciated with test homogeneity?
* A. Kuder-Richardson
B. Test-Retest
C. Equivalent-Tests
D. Split-halves

AC15 Which of the following estimates of reliability is most clearly associated with the Spearman-Brown Prophecy Formula?
A. Kuder-Richardson
B. Test-Retest
C. Equivalent-Tests
* D. Split-halves

AC16 The average score that a person would make over repeated trials on the same test is his:
A. reliability
* B. true score
C. obtained score
D. error variance

AC17 The standard deviation of the distribution of error scores is called the:
A. reliability of the test
B. reliability error
C. standard error of estimate
* D. standard error of measurement

AC18 The extent to which a test truly represents the area of knowledge under consideration is its:
A. face validity
* B. content validity
C. criterion validity
D. construct validity

AC19 The type of validity that is particularly relevant when evaluating personality measures is:
A. content validity
* B. construct validity
C. criterion-related validity
D. face validity

AC20 If raw scores on a test are normally distributed, the greatest difference in raw score points will be between percentile ranks:
* A. 1 and 5
B. 25 and 30
C. 50 and 55
D. 90 and 95

AC21 Ernest scored at the 99th percentile on the entrance test at Cascade College. The best interpretation of his score is:
A. He should obtain higher grades than 99 percent of the students.
B. He should obtain high grades with relatively little effort.
C. His score is comparable to an IQ of about 130.
* D. He scored higher than 99 percent of the students taking the test.

AC22 Which of the following scores is expressed in raw score units?
A. stanines.
* B. percentile points
C. normalized standard scores
D. percentile ranks

AC23 The mean and standard deviation of the distribution of z scores are, respectively:
* A. 0 and 1.
B. 10 and 3
C. 50 and 10
D. 100 and 15.

AC24 The first question to be asked when evaluating a standardized achievement test is:
A. What is the editorial quality of the test?
* B. What does the test measure?
C. How reliable is the test?
D. Are equivalent forms of the test available?

AC25 If a test is valid, then the test:
A. scores must be normally distributed
* B. must be reliable
C. must be relatively long
D. must have national norms

BC01 Objectives have not been defined for which of the following domains?
* A. psychological
B. cognitive
C. psychomotor
D. none of the above

BC02 Which of the following terms is least acceptable for instructional objectives which are to be measured through multiple-choice test items?
A. recognize
* B. recall
C. identify
D. differentiate

BC03 Which of the following is most correct? Instructional objectives should:
A. end with an active verb
B. be stated in terms of teacher behavior
C. start with an active verb
* D. be stated in terms of student behavior

BC04 Which of the following is not correct? "Standardized tests provide a standard for:
A. timing
B. scoring
* C. precision
D. administration

BC05 The word "criterion' in 'criterion-referenced test' usually refers to:
* A. a set of objectives
B. criterion validity
C. students' scores on a previous test
D. a standardized achievement test

BC06 The assignment of numerals to objects or events according to rules is a definition of:
A. evaluation
B. testing
* C. measurement
D. statistics

BC07 In educational measurement, we usually assume that the scale of measurement is:
A. nominal
B. ordinal
* C. interval
D. ratio

BC08 The number of students who get an item correct divided by the total number of students is called:
* A. difficulty level
B. error rate
C. theoretical difficulty level
D. theoretical error rate

BC09 A statistic used to show how sharply an item differentiates between students who score high on a test and students who score low is called a:
* A. discrimination index
B. standard deviation
C. correlation coefficient
D. standard error of measurement

BC10 Which of the following is not a possible value of the standard deviation of a set of scores?
A. 10.01
B. 0.00
* C. -0.01
D. 201.00

BC11 The median is also called the:
A. variance
* B. 50th percentile
C. semi-interquartile range
D. mode

BC12 Which of the following is not a possible value of the Pearson product-moment correlation coefficient?
A. -0.0001
B. -1.0000
C. 0.9999
* D. 1.0001

BC13 If test scores are distributed normally, what percent of the scores will exceed a score falling one standard deviation above the mean?
* A. 16%
B. 34%
C. 68%
D. 84%

BC14 Which of the following estimates of reliability is most clearly associated with internal consistency?
A. split-halves
B. equivalent-tests
* C. Kuder-Richardson
D. test-retest

BC15 Which of the following estimates of reliability typically employs the Spearman-Brown Prophecey Formula?
A. Kuder-Richardson
B. parallel-tests
* C. Split-halves
D. coefficient alpha

BC16 The average score that a person would receive over repeated administrations of the same test is his:
A. theoretical score
B. observed score
* C. true score
D. error score

BC17 The standard deviation of the distribution of error scores is called the:
A. standard deviation
B. reliability of the test
C. error deviation
* D. standard error of measurement

BC18 The type of validity most appropriate for achievement tests is:
A. face validity
* B. content validity
C. construct validity
D. criterion validity

BC19 The type of validity most clearly associated with theories of personality is:
A. face validity
B. content validity
C. criterion validity
* D. construct validity

BC20 If raw scores are normally distributed, the greatest difference in raw score points will be between percentile ranks:
* A. 1 and 5
B. 48 and 52
C. 70 and 74
D. 94 and 98

BC21 Jerry scored at the 75th percentile on the SAT-Mathematics test. The best interpretation of his score is:
* A. he scored higher than 75 percent of the students who took the test.
B. he scored higher than 25 percent of the students who took the test.
C. his IQ is above average
D. his IQ is below average

BC22 Which of the following scores is expressed in raw score units?
A. T-scores
B. Z-scores
* C. percentile points
D. percentile ranks

BC23 The mean and standard deviation of z-scores are respectively:
A. 5 and 2
B. 50 and 10
C. 100 and 16
* D. none of the above

BC24 The most important aspect of a standardized achievement test is its:
* A. content
B. reliability
C. editorial quality
D. cost

BC25 If a test is valid, then the test must be
A. relatively long
* B. reliable
C. standardized
D. normally distributed

ZC26 Which is the best example of a free-response test?
* A. an essay test
B. a matching test
C. a multiple-choice test
D. a short-answer test

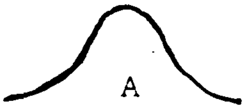
ZC27 In a pure power test, test takers would not differ in the:
* A. number of items attempted
B. number of items answered correctly
C. percent of items answered correctly
D. time taken to complete the test

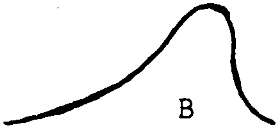
ZC28 Which of the following types of tests is least apt to be used in order to rank order students?
A. norm-referenced test
B. standardized test
* C. criterion-referenced test
D. non-standardized test

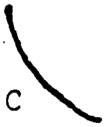
ZC29 In educational measurement, we usually tacitly assume that the underlying scale of measurement is:
A. nominal
B. ordinal
* C. interval
D. ratio

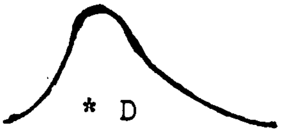
ZC30 The percentage of students that we expect will get an item wrong if everybody guesses blindly is called:
A. difficulty level
B. error rate
C. theoretical difficulty level
* D. theoretical error rate

ZC31 Which of the following is a positively skewed distribution?


A


B


C


* D

ZC32 In which of the following distributions do the mean, median, and mode always coincide?
* A. Normal
B. Positively skewed
C. Bimodal
D. J-shaped

ZC33 Which of the following is not an index of the dispersion of a set of test scores?

* A. the mean
B. the standard deviation
C. the variance
D. the range

ZC34 An index measuring the degree of relationship between two different measures for a group of individuals is a:

* A. correlation coefficient
B. standard error of measurement
C. discrimination index
D. standard deviation

ZC35 If high values of X are associated with low values of Y, and low values of X are associated with high values of Y, then:

A. $r_{xy} = 0$

B. $r_{xy} > 0$

* C. $r_{xy} < 0$

D. r_{xy} is undefined

ZC36 Which of the following statements is true?

A. Different methods of computing a reliability coefficient will give the same result.
B. Very hard tests generally have higher reliabilities than very easy tests.
* C. A longer test is generally more reliable than a shorter one.
D. Older tests, which have been used more, are generally more reliable than newer ones.

ZC37 Which of the following estimates of reliability takes into account the most sources of variation?

A. Test-Retest without time interval intervening.
B. Test-Retest with time interval intervening.
C. Equivalent-Tests without time interval intervening.
* D. Equivalent-Tests with time interval intervening.

ZC38 The reliability of a test refers to:

A. how accurately the test measures the trait it is designed to measure.
* B. the precision with which the test measures whatever it measures.
C. how accurately the test categorizes people into defined groups.
D. how much faith you can put in the test scores.

ZC39 Mary has taken an intelligence test during each of the last three years. Her scores on successive testings were 121, 118, and 114. The most reasonable explanation of these results is:

A. Her scores are dropping as competition gets rougher at older age levels.

B. A personal problem is probably interfering with her performance.

C. The test used is not good as it does not measure consistently.

* D. The scores are within the range expected on repeated testings.

ZC40 Which of the following is most useful for estimating a person's true score?

A. the reliability of the test.

* B. the standard error of measurement

C. the mean of the test

D. the standard deviation of the test

ZC41 Scores on a final exam in introductory psychology are correlated with scores on a well-known norm-referenced test in psychology. The resulting correlation coefficient is evidence of the final exam's:

A. face validity

B. content validity

* C. criterion validity

D. construct validity

ZC42 Which of the following is <u>not</u> an essential requirement of a criterion measure?

A. Measure an important component of the task

B. measures reliability

* C. measures more than one behavior

D. is free from bias

ZC43 The weakest link in most validity studies is the:

A. predictors

* B. criterion

C. sample

D. validation technique

ZC44 The correlations between predictors and performance in an auto mechanics lab are: compulsivity, +.26; mechanical comprehension, +.33; intelligence, +.05; English grades, -.43. The best predictor of performance in the lab is:

A. compulsivity scores

B. mechanical comprehension

C. intelligence

* D. English grades

ZC45 If, when predicting college grades from a college admissions test, r_{xy} = 50, we can say that:

* A. 25 percent of the variance in grades is predictable from the test scores.
B. 50 percent of the variance in grades is predictable from the test scores.
C. Using the test will reduce prediction errors by 50 percent.
D. Predicting from the test will be 25 percent better than chance predictions.

ZC46 A test has a mean of 45 and a standard deviation of 10 points. If the distribution of scores is approximately normal, the range of scores in a class of 50 students would be from approximately:
A. 40 to 50
B. 35 to 55
C. 25 to 65
* D. 15 to 75

ZC47 Norms are most useful for:
A. selecting the best qualified workers
* B. comparing a person to his immediate competitors.
C. studying the extent of individual differences.
C. computing the validity of a test.

ZC48 Which of the following are least useful for evaluating college students?
* A. age norms
B. grade norms
C. percentile norms
D. standard scores

ZC49 Consistency of measurement is often called:
* A. reliability
B. validity
C. variance
D. skewness

ZC50 In preparing students to take standardized achievement batteries, the teacher should:
A. drill the students on the material to be covered on the test.
* B. briefly explain the nature and purpose of the test.
C. keep reminding the students how important the test will be.
D. say nothing in advance of the test so students will not become anxious.