A Descriptive Exploratory Analysis of Some Issues in Criterion-Referenced Measurement with Possible Application to a Diagnostic-Prescriptive System for Developing Measurement Competency for Prospective Teachers

Allen Norton Shub
Loyola University Chicago

Follow this and additional works at: https://ecommons.luc.edu/luc_diss

Part of the Education Commons

Recommended Citation
https://ecommons.luc.edu/luc_diss/1699

This Dissertation is brought to you for free and open access by the Theses and Dissertations at Loyola eCommons. It has been accepted for inclusion in Dissertations by an authorized administrator of Loyola eCommons. For more information, please contact ecommons@luc.edu.

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License.
Copyright © 1977 Allen Norton Shub
A DESCRIPTIVE EXPLORATORY ANALYSIS OF SOME ISSUES IN CRITERION-REFERENCED MEASUREMENT WITH POSSIBLE APPLICATION TO A DIAGNOSTIC-PRESCRIPTIVE SYSTEM FOR DEVELOPING MEASUREMENT COMPETENCY FOR PROSPECTIVE TEACHERS

by

Allen N. Shub

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

June 1977

© Allen N. Shub 1977

All rights reserved
ACKNOWLEDGEMENTS

The author wishes to express his gratitude to the members of his doctoral committee chaired by Dr. Samuel T. Mayo and consisting of Dr. Jack A. Kavanagh and Dr. Ernest I. Proulx for direction and guidance in the preparation of this dissertation. The author also wishes to thank Dr. Ronald R. Morgan and his two classes in tests and measurements for their participation in the field tryout. A special debt is owed to my wife, Susan Barasch Shub, for her support, encouragement, assistance, and sacrifices during the preparation of this doctoral dissertation. Finally, the author wishes to express his appreciation to Mr. Richard R. Meredith for the typing of the final manuscript.
LIFE

The author, Allen N. Shub, is the son of Seymour V. Shub and Estelle M. (Dreebin) Shub. He was born June 7, 1943, in Chicago, Illinois.

His elementary and secondary education was obtained in Chicago schools. He graduated from South Shore High School in 1961.

In September 1961 he entered the University of Illinois (Urbana) as an Edmund J. James Scholar and in June 1965 received the degree of Bachelor of Science with a major in psychology and minors in chemistry and zoology.

In September 1965 he was granted a research assistantship in psychology at Loyola University of Chicago. In February 1968 he was awarded the degree of Master of Arts in social psychology.

In September 1967 he was granted a research assistantship and in September 1968 a National Institute of Health traineeship in quantitative psychology at the L. L. Thurstone Psychometric Laboratory of the University of North Carolina at Chapel Hill.

In July 1968 he married Susan C. Barasch.

Since September 1969 he has been employed by Science Research Associates, Inc. (SRA), Chicago, a subsidiary of IBM.

In June 1970 he entered the School of Education at Loyola University of Chicago to pursue doctoral studies under the direction of Dr. Samuel T. Mayo, Professor, Department of Foundations of Education.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>LIFE</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>CONTENTS OF APPENDICES</td>
<td>vii</td>
</tr>
<tr>
<td><strong>Chapter</strong></td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. HISTORICAL BACKGROUND</td>
<td>13</td>
</tr>
<tr>
<td>Early History</td>
<td>14</td>
</tr>
<tr>
<td>Assessment of Academic Achievement</td>
<td>16</td>
</tr>
<tr>
<td>Washburne's Winnetka Plan</td>
<td>18</td>
</tr>
<tr>
<td>Morrison's University of Chicago Laboratory</td>
<td>20</td>
</tr>
<tr>
<td>School Approach</td>
<td>22</td>
</tr>
<tr>
<td>Other American Roots</td>
<td></td>
</tr>
<tr>
<td>III. TECHNICAL ISSUES</td>
<td>25</td>
</tr>
<tr>
<td>Problems and Questions</td>
<td>25</td>
</tr>
<tr>
<td>Behavioral Objectives</td>
<td>31</td>
</tr>
<tr>
<td>Writing Objectives</td>
<td>33</td>
</tr>
<tr>
<td>Selecting Objectives</td>
<td>37</td>
</tr>
<tr>
<td>Test Construction</td>
<td>41</td>
</tr>
<tr>
<td>Item Writing</td>
<td>41</td>
</tr>
<tr>
<td>Number of Items Per Objective</td>
<td>48</td>
</tr>
<tr>
<td>Item Analysis</td>
<td>51</td>
</tr>
<tr>
<td>Reliability</td>
<td>58</td>
</tr>
<tr>
<td>Validity</td>
<td>65</td>
</tr>
<tr>
<td>Mastery Criterion</td>
<td>72</td>
</tr>
<tr>
<td>Norm-Referenced Criterion-Referenced Measures</td>
<td>80</td>
</tr>
<tr>
<td>Construction of a Diagnostic-Prescriptive System</td>
<td>83</td>
</tr>
<tr>
<td>for Developing Measurement Competency for Prospective Teachers.</td>
<td></td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Characteristics of Norm-Referenced and Criterion-Referenced Tests</td>
<td>7</td>
</tr>
<tr>
<td>2. Probability of Attaining Mastery by Chance Alone</td>
<td>79</td>
</tr>
<tr>
<td>3. Item Numbers of Measurement Competency Test Corresponding to Content Categories</td>
<td>88</td>
</tr>
<tr>
<td>4. Means and Standard Deviations for Measurement Competency Test (Form A): Experimental Group</td>
<td>94</td>
</tr>
<tr>
<td>5. Means and Standard Deviations for Measurement Competency Test (Form A): Control Group</td>
<td>96</td>
</tr>
<tr>
<td>6. Analysis of Covariance of MCT Posttest Scores</td>
<td>97</td>
</tr>
<tr>
<td>8. Analysis of Covariance of MCT Posttest Scores: Eliminating Four Outliers</td>
<td>100</td>
</tr>
</tbody>
</table>
## CONTENTS OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Checklist of Measurement Competencies</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>(Reprinted, with permission, from Mayo, 1967)</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Measurement Competency Test</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>(Reprinted, with permission, from Mayo, 1967)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>I. Form A</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td>II. Form B</td>
<td>143</td>
</tr>
<tr>
<td>C</td>
<td>Diagnostic Tests</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>I. Standardized Tests</td>
<td>152</td>
</tr>
<tr>
<td></td>
<td>II. Construction and Evaluation of Classroom Tests</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>III. Uses of Measurement and Evaluation</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>IV. Statistical Concepts</td>
<td>179</td>
</tr>
<tr>
<td>D</td>
<td>Answer Keys</td>
<td>193</td>
</tr>
<tr>
<td></td>
<td>I. Standardized Tests</td>
<td>194</td>
</tr>
<tr>
<td></td>
<td>II. Construction and Evaluation of Classroom Tests</td>
<td>196</td>
</tr>
<tr>
<td></td>
<td>III. Uses of Measurement and Evaluation</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td>IV. Statistical Concepts</td>
<td>200</td>
</tr>
<tr>
<td>E</td>
<td>Prescription Sheets for Stanley &amp; Hopkins (1972)</td>
<td>204</td>
</tr>
<tr>
<td></td>
<td>I. Standardized Tests</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>II. Construction and Evaluation of Classroom Tests</td>
<td>206</td>
</tr>
<tr>
<td></td>
<td>III. Uses of Measurement and Evaluation</td>
<td>207</td>
</tr>
<tr>
<td></td>
<td>IV. Statistical Concepts</td>
<td>208</td>
</tr>
<tr>
<td>F</td>
<td>Instructor's Guide</td>
<td>212</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

Accountability. Guaranteed performance. Pay for performance. Relevance. Individualized instruction. Test scores. The issues reflected by these terms have become burning ones in recent years as parents and educators alike are demanding that the educational process be increasingly responsive to the individual needs of children. How do we determine these needs? The criterion-referenced measurement point of view suggests one solution.

What is criterion-referenced measurement? From all the recent discussions in journals and with educational test publishers now emphasizing criterion-referenced tests, one would think they are something new. Actually, they have been around for a long time, although not always by that name. Robert Glaser (1963) is generally credited with introducing the term criterion-referenced measurement and distinguishing it from norm-referenced measurement. Let's examine the distinction made between these two types of tests.

Tests can be classified in many ways, such as by content (e.g., history, English, mathematics, etc.), by purpose (e.g., achievement, intelligence, aptitude, interest, personality), by item type (e.g., essay, multiple-choice, etc.), or by type of comparison or score reporting (e.g.,
norm-referenced, criterion-referenced). For this discussion, they will be defined according to the last distinction.

A norm-referenced test gives us information about what the student has learned or can do compared to what others have learned or can do. Interpretation of scores is relative, that is, in relation to a "norm group," traditionally defined as a representative sample from an appropriate population.

The term norm-referenced measurement can be applied to traditional standardized achievement tests, where scores typically reported are grade equivalents and/or percentiles. A grade equivalent is the grade placement of pupils in the norm sample for which the obtained raw score is the median level of performance. For example, if a pupil obtains a grade-equivalent score of 7.3, this means that his raw score is the same on this test as the median score made by pupils in the third month of the seventh grade.

A percentile rank is the percentage of pupils at a given grade level in the norm sample who obtained scores lower than the corresponding raw score. For example, a percentile rank of 55 for a fifth grader would represent a raw score higher than 55% of the scores of fifth graders on this test in the norm group.

Norm-referenced tests have been and remain useful tools in the educational process. As Farr and Roser (1974) reported:

The development and use of standardized tests have resulted from demands for useful sources of information: information for planning instruction, for estimating students' growth, and for assessing a school district's success in achieving stated goals. (p. 592)
Harsh (1974), in discussing the development of norm-referenced tests, pointed out:

The construction of the national, standardized NRT was based on surveys of contents, materials and anticipated outcomes of schools in every region. Courses of study, curriculum guides, textbooks, instructional materials and educators' definitions were compiled and analyzed to identify contents with the highest common incidence. Items of these nationally standardized tests were designed as surveys of skills and knowledges generally common to many or most educational programs. (p. 3)

Harsh further noted that the standardized norm-referenced test has an "imperfect and incomplete congruence to any particular school program" (p. 3). Or, as Vernon S. Larsen (1971b) has pointed out, tongue-in-cheek: "A survey test is carefully designed to be equally unfair to all curricula."

In contrast to a norm-referenced test, a criterion-referenced test gives us specific information as to what the student has learned versus what he was expected to have learned. Several definitions of criterion-referenced tests appear in the literature. Probably the definition most widely cited is that of Glaser and Nitko (1971):

A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards. Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual. Measurements are taken on representative samples of tasks drawn from this domain, and such measurements are referenced directly to this domain for each individual. (p. 653)

Wang (1969) described a criterion-referenced test as "an achievement test developed to assess the presence or absence of a specific criterion behavior described in an instructional objective."
According to Harris and Stewart (1971):

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.

Burns (1972) defined a criterion-referenced test as "a measure of the degree of the effectiveness of the interaction between the elements of instruction, the strategy presented for learning and the learning style and ability of the learner" (p. 42).

According to Popham and Husek (1969):

Criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion, i.e., a performance standard. It is because the individual is compared with some established criterion, rather than other individuals, that these measures are described as criterion-referenced. (p. 2)

Nitko (1970) described a criterion-referenced test as "one that is deliberately constructed to give scores that tell us what kinds of behaviors individuals with those scores can demonstrate" (p. 38).

Alkin (1974, p. 4) noted that common characteristics of criterion-referenced tests are that they are organized around behavioral objectives and they provide assessment with respect to predefined performance criteria.

Other researchers have made further distinctions among various types of criterion-referenced measures, e.g., domain-referenced tests (cf. Denham, 1975; Hively, Maxwell, Rabehl, Sension, & Lundin, 1973; Sanders & Murray, 1976).
The term criterion-referenced measurement can be applied both to mastery and to diagnostic tests. A **mastery test** is not intended to indicate how much a student has achieved relative to other students, but to demonstrate his or her strengths or weaknesses in a given area. A **diagnostic test**, on the other hand, is used to locate specific areas of weakness and to determine the extent of these. In effect, mastery and diagnostic measurement can be considered in the same context; it is the purpose for which the test is used which determines whether it is one or the other. Vernon S. Larsen's (1971a) tongue-in-cheek distinction between the two will further elaborate on this point: "A mastery test is a diagnostic test given too late to do any good!"

Is there really a difference between norm-referenced and criterion-referenced tests? Greco (1974) presented an argument that the two types of tests are not so different as researchers in the area have seemed to have us believe and that "it appears that the recent literature relating to criterion-referenced tests has little relevance for the individual teacher" (p. 25). Other investigators seem to see a symbiotic or complementary relationship between the two types of tests. According to Guzaitis (1973):

The emergence of criterion-referenced tests has been perceived as a threat to norm-referenced testing, but it need not be viewed in this way since it seems unlikely that parents, teachers, and administrators will suddenly lose the desire to know where their students rank in relation to others. Hopefully, they will no longer try to force the norm-referenced test to do a double duty it was not built for.
Harsh (1974) expanded on the theme of a perceived threat from criterion-referenced tests and argued:

... By design, the NRT and CRT are conceived with different frames of reference. They are not totally exclusive of each other, but they do direct attention at different uses and inferences for interpretation and decision making. Moreover, we commend the notion that rather than viewing NRT and CRT as adversaries seeking victory over each other, their combined contributions allow a more detailed and comprehensive means of assessing and evaluating outcomes of an educational program. (p. 1)

Mayo (1970) has pointed out that a criterion-referenced test score should not be considered relative but absolute; interpretation of the test score can be made in terms of describing the specific behaviors which a student can exhibit. In other words, where a norm-referenced test score determines a relative ranking of students, a criterion-referenced test score identifies specific accomplishments or weaknesses.

Boehm (1973, p. 119) presented an excellent summary table of characteristics and distinctions of norm-referenced tests and criterion-referenced tests. (See Table 1.)

If criterion-referenced measurement has such a rich history, as evidenced by the material presented in Chapter II, one could reasonably ask why has the emphasis on this topic among educators and educational test publishers been comparatively recent? As Hawes (1973) noted:

Local schoolmen started it, state educators picked it up, then test publishers jumped on the bandwagon. Together they've made criterion-referenced testing easily the fastest-growing new technique for evaluating school achievement today.

Why the wildfire interest in this novel testing method? Accountability, of course. Criterion-referenced testing, say its proponents,
<table>
<thead>
<tr>
<th></th>
<th>Norm-Referenced</th>
<th>Criterion-Referenced</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General Purpose</strong></td>
<td>To make comparisons among</td>
<td>To determine how an individual functions relative to a criterion</td>
</tr>
<tr>
<td></td>
<td>individuals</td>
<td></td>
</tr>
<tr>
<td></td>
<td>To make decisions about</td>
<td>To program specifically for the individual</td>
</tr>
<tr>
<td></td>
<td>placement in programs in which</td>
<td></td>
</tr>
<tr>
<td></td>
<td>only limited numbers of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>individuals can be accepted</td>
<td></td>
</tr>
<tr>
<td></td>
<td>To determine for whom a</td>
<td>To determine whether an instructional program &quot;works&quot; in developing criterion behaviors</td>
</tr>
<tr>
<td></td>
<td>program &quot;works&quot;</td>
<td></td>
</tr>
<tr>
<td><strong>Item Types</strong></td>
<td>Items must discriminate</td>
<td>Items must correspond to criterion levels</td>
</tr>
<tr>
<td></td>
<td>among individuals</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Items all subjects pass or</td>
<td>Items must provide explicit information about what an individual can or cannot do</td>
</tr>
<tr>
<td></td>
<td>all fail eliminated</td>
<td></td>
</tr>
<tr>
<td><strong>Content</strong></td>
<td>Content may or may not match</td>
<td>Content must match classroom objectives which have been behaviorally defined beforehand</td>
</tr>
<tr>
<td></td>
<td>particular classroom goals</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sampling is made from the</td>
<td>Criterion levels can be set at each content level of a program and must specify minimal levels of competence</td>
</tr>
<tr>
<td></td>
<td>larger task domain</td>
<td></td>
</tr>
<tr>
<td><strong>Scores</strong></td>
<td>Variability among scores is</td>
<td>Variability is irrelevant</td>
</tr>
<tr>
<td></td>
<td>essential</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scores can mask what an</td>
<td>Scores must reflect (not mask) what an individual can or cannot do</td>
</tr>
<tr>
<td></td>
<td>individual can do but provide</td>
<td></td>
</tr>
<tr>
<td></td>
<td>indication of his relative</td>
<td></td>
</tr>
<tr>
<td></td>
<td>standing</td>
<td></td>
</tr>
<tr>
<td><strong>Type of Ranking</strong></td>
<td>Use of age and grade norms;</td>
<td>Percentage passing a criterion level; pass/fail information on each item</td>
</tr>
<tr>
<td></td>
<td>percentiles; standard scores</td>
<td></td>
</tr>
</tbody>
</table>

allows a school or district to measure and report accomplishments well beyond the scope of traditional tests. (p. 35)

Hawes's comments on accountability were echoed by Farr and Roser (1974):

But perhaps the strongest impetus for more testing [criterion-referenced tests] has resulted from the advance of the accountability concept. Accountability has been simply defined as holding the educational profession responsible for the growth of the children in its control. Usually this has been interpreted to mean that tests should be given to the children to find out how well the educators have done their jobs. (p. 595)

Airasian and Madaus (1972) noted two undesirable effects of grading student performance on a norm-referenced basis:

1. They [norm-referenced grading practices] have given credence to the notion that for success or achievement to mean anything, there must be a reference group of nonattainers. The rewards system engendered by norm-referenced grading insures "winners" and "losers" in the achievement race.

2. Norm-referenced practices have led to a discrepancy between the rewards system (i.e., grades) and the actual performance of students. (p. 1)

Airasian and Madaus then noted four trends leading to the growing interest in criterion-referenced measures since the late 1960s: (a) growing criticism of testing, especially standardized achievement and ability tests; (b) growing controversy concerning grades; (c) growth of the instructional technology movement; and (d) growing belief that "all or at least most students can learn, benefit from, or be helped to achieve competency in most subject areas" (p. 3).

Hunt (1975) believes that norm-referenced tests have "sabotaged the educational process" (p. 343). He gave three ways in which this
9

is so:

1. . . . The idea that the IQ provides a measure of future potential as well as past achievement has tended to produce in teachers' expectations for their pupils that damped their ingenuity in instruction . . . [and expectations which], thereby, became self-fulfilling prophecies. . . .

2. . . . Intelligence tests served as the norm-referenced models for the testing of educational achievement in general. The emphasis on interpersonal comparison in this model has distracted both pupil and teacher from the specific goals of each individual task in the teaching-learning enterprise. It has also served to diminish the self-confidence of a preponderant majority of pupils; for, if excellence is defined in terms of membership in the top ten percent, then ninety percent must fail to achieve excellence.

3. . . . This extension of the norm-referenced model from tests of intelligence to tests of achievement has separated testing from the teaching-learning enterprise. As a consequence, the children in schools are put in lock-step with given curricula badly matched to their individual abilities and interests. They are then examined with tests only distantly relevant to their specific learning tasks. (pp. 342-343)

Finally, Farr and Roser (1974) noted that controversy over the use of standardized tests in schools has occurred because of reasons such as the following:

1. Tests are sometimes administered when there is no clearly stated purpose for administering them. . . .

2. Test results are often viewed as the sole criteria on which to judge the success of a program. . . .

3. Specific tests which assess specific goals are often used to assess the achievement of all goals. . . .

4. Test results are released to the news media and the public without accompanying information. . . .

5. Test results purportedly collected to aid in classifying children and in planning educational programs have been used to rigorously label and inflexibly track students. . . . (p. 593)

In general, it appears that the new emphasis in the field for criterion-referenced measures is the direct result of a growing concern that
standardized achievement tests are not providing relevant information to teachers, students, and administrators. Typically the teacher learns from a standardized achievement test that one student falls, say, at the 25th percentile in mathematics computation, and another falls at the 58th percentile. This tells us very little about what the student does or does not know. Standardized achievement testing certainly has its important uses in educational assessment, but pinpointing specific weaknesses and strengths is not one of them. It would be valuable for the teacher and student to know which of the many skills underlying mathematics computation (and concepts) the student is weak in: Is he or she weak in "adding two 2-digit numbers, regrouping tens and hundreds" or in "dividing a mixed number by a fraction with numerator greater than one, with cancellation possible"? Criterion-referenced diagnostic tests can quickly and easily uncover these and other weaknesses where they exist. And a diagnostic-prescriptive system can in addition prescribe appropriate remediation for each learning objective.

Brazziel (1972) listed six advantages of criterion-referenced tests:

1. Permit direct interpretation of progress in terms of specified behavioral objectives.
2. Facilitate individualized instruction.
3. Eliminate a situation where half of American schoolchildren must always be below the median. . . .
4. Are usually short summative tests which enable teachers to check on student progress at regular intervals.
5. Eliminate pressures on teachers to "teach to the test" in order to have children make a good showing.
6. Enable teachers to compile a comprehensive record of the child's development. (pp. 52-53)

Criterion-referenced measurement should not be seen as the remedy for all the problems of norm-referenced measurement, however. For example, Knipe and Krahmer (1973) view criterion-referenced testing "as a meaningful contribution to education, not as a panacea to all the ills of educational testing and measurement."

And criterion-referenced measurement, though an old concept, is a relatively young science and has its unique measurement problems. For example, how should criterion-referenced tests be constructed? How should the "goodness" of the test items be evaluated? How many test items are needed to measure an objective and to certify attainment of "mastery"? How should reliability and validity of a criterion-referenced test be assessed? The purpose of this dissertation, then, is to identify and clarify the difficulties in the field of criterion-referenced measurement, restricted to "paper and pencil" tests. The dissertation presents, reviews, and discusses the major psychometric issues and controversies in criterion-referenced measurement with a view toward recommending some resolutions. A secondary purpose is to apply the state of the art to producing a diagnostic-prescriptive system for developing measurement competency for prospective teachers. An important significance of this dissertation is the manner in which the criterion-referenced issues of behavioral objectives, test construction, reliability, validity, mastery
criterion, and norm-referenced criterion-referenced measures are compiled and analyzed in one work.

Chapter II presents an historical background to criterion-referenced measurement. Chapter III presents a discussion and analysis of some important technical issues currently confronting the field and describes application of criterion-referenced procedures to the construction of a diagnostic-prescriptive system for developing measurement competency for prospective teachers. Chapter IV describes the field tryout and results of the system. Chapter V presents a discussion of the results and presents possible resolutions of some of the technical issues in criterion-referenced measurement.
CHAPTER II

HISTORICAL BACKGROUND

Chapter I presented an introduction to and definitions of norm-referenced measurement and criterion-referenced measurement, including diagnosis and mastery, and discussed rationale, purposes, and advantages and disadvantages of norm-referenced and criterion-referenced measures. Chapter II presents an historical background to criterion-referenced measurement and demonstrates that the concept of criterion-referenced measurement has been with us as early as the 23rd Century B.C.

Researchers often trace attempts at mastery learning and diagnosis in education as far back as the 1920s: Carleton Washburne's Winnetka Plan and Henry C. Morrison's University of Chicago Laboratory School strategy. These approaches both involved defining mastery in terms of specific educational objectives, providing for measures of diagnosis and mastery, and providing for appropriate supplemental instruction or remediation (Block, 1971, p. 3).

Actually, we can trace the history of criterion-referenced measurement back much further—-to biblical days for example, where criterion-referenced measurement was done in a do-or-die manner. It is told that the Gileadites in the 12th Century B.C. devised a test to determine whether
strangers were enemy Ephraimites or friendly Gileadites: "Say now shibboleth," the strangers were told. Those who could pronounce the word correctly were determined to be Gileadites and were allowed to pass; those who said "sibboleth" were deemed Ephraimites (for the Ephraimites were unable to pronounce "shibboleth") and hence were slain. "And there fell at that time of the Ephraimites forty and two thousand" (Judges 12: 6). It is not known how many passed this crucial test. (Sometimes this writer wonders whether the origin of the term password comes from this biblical selection, for those who were able to say the word were allowed to pass and those who were unable to say the word were thrown over the pass.)

Early History

If constructed to measure mastery versus nonmastery of a particular subject matter, civil service examinations can be considered to be practical, criterion-referenced tests. The ancient Chinese had an elaborate civil service system by the end of the 2nd Century B.C. (Han Dynasty), although the beginnings of the Chinese civil service examinations appear to be placed as early as the 23rd Century B.C. with the Emperor Shun (DuBois, 1965, p. 4; DuBois, 1970, p. 3; Ebel, 1972, p. 5). The early system was later refined during the Sung Dynasty (960-1279 A.D.). The Sung system involved the successive attrition of candidates through written tests on three levels--"budding geniuses," "promoted scholars," and "ready for office" (DuBois, 1965, p. 5) (only one of every hundred candidates was said to
have successfully passed)--with anonymity guaranteed by clerks recopying the exam with candidates identified only by number. Three readers were required to read each candidate's paper. In addition, the system provided for annual merit ratings and promotions from within the ranks (Encyclopaedia Britannica, 1971, vol. 21, p. 425).

Around 2,400 years after the beginnings of the Chinese civil service system, and thousands of miles away, the Roman system was being founded by Augustus around the 1st Century A.D. (Encyclopaedia Britannica, 1971, vol. 19, p. 527).

Furthermore, in ancient Greece, we are told, every Spartan boy had a series of tests through which he had to pass in demonstrating his attainment of the required skills of manhood. And, in Athens, young scholars were subjected to Socratic inquiry to demonstrate their competence (Chauncey & Dobbin, 1963, p. 1).

The Chinese civil service system influenced the French and the British civil service examinations systems in the 19th Century; and the British experience led to the use of the civil service examination by the United States (DuBois, 1970, pp. 5-6). The U.S. Civil Service Act of January 16, 1883, establishing competitive tests for entry into government jobs, required that:

... such examinations shall be practical in their character, and so far as may be shall relate to those matters which will fairly test the relative capacity and fitness of the persons examined to discharge the duties of the service into which they seek to be appointed. (DuBois, 1970, p. 6)
During the early years of the U.S. Civil Service Commission, technical procedures included:

1. Study of characteristics of individuals performing successfully in positions to be filled by examination.
2. Development of examination questions intended to measure these characteristics.
3. Administration of examinations under conditions intended to give each candidate an equal opportunity to succeed.
4. The use of a system of examination numbers to conceal the identity of candidates from examiners reading the tests.
5. The use of a carefully drawn point system in the scoring of tests so that the final score would not vary appreciably with the examiner assigned to read the test.
6. Ranking the candidates in order of grade as a step in determining eligibility for appointment. (DuBois, 1970, p. 7)

That similarities should be noted between the U.S. Civil Service Commission and the Chinese civil service system should not be surprising. The Chinese system heavily influenced the British system which in turn influenced the United States system.

Assessment of Academic Achievement

Another influence of today's emphasis and interest in criterion-referenced measurement can be traced back to the assessment of academic achievement. Originally developed to ascertain qualifications for academic degrees, these examinations were for hundreds of years exclusively oral examinations. The first oral examination may well have been the ones administered by the University of Bologna for law candidates in 1219. Oral examinations were also found in 1257 at what is now known as the Sorbonne and in 1441 at Louvain University. Recall that the Chinese civil service
system had three levels ("budding geniuses," "promoted scholars," and "ready for office"); the competitive examinations at Louvain ranked candidates into four classes: "honor men," "satisfactory," "charity passes," and "failures" (DuBois, 1970, p. 8).

Written examinations were apparently pioneered by the Jesuit order for placement and post-instruction evaluation. In 1599, the definitive "Ratio Studiorum" was published to provide specific procedures for the conduct of examinations (DuBois, 1970, pp. 8-9).

Oral examinations for university degrees appear to have been used extensively in England in 1636 at Oxford University. By 1803, written examinations were used at Oxford—earlier at Cambridge University; and printed examinations were used in 1828 (DuBois, 1970, p. 10). In 1836, the University of London was chartered solely to examine candidates for degrees; for years it had no instructional program at all (DuBois, 1970, p. 10).

Two American educators, Horace Mann and Emerson E. White, argued for the use of written examinations over oral examinations. Mann, Secretary of the Massachusetts Board of Education, showed in 1845 how oral examinations were lacking in what today would be called validity, reliability, and usability (Ross & Stanley, 1954, p. 29). The advantages of written examinations included:

1. More evidence could be obtained of the achievements of each pupil.
2. A written record of these achievements would be produced.
3. Each pupil would be asked the same questions; thus all would be treated alike.
4. There would be less possibility of favoritism for or bias against particular pupils or teachers. (Ebel, 1972, p. 7)
White, in 1886, argued that the written test

... is more impartial than the oral test, since it gives all the pupils the same tests and an equal opportunity to meet them; its results are more tangible and reliable; it discloses more accurately the comparative progress of the different pupils, information of value to the teacher; it reveals more clearly defects in teaching and study, and thus assists in their correction; it emphasizes more distinctly the importance of accuracy and fullness in the expression of knowledge; it reveals more fully than the ordinary language exercise the ability of the pupil to write correctly when his attention is directed to the thought of the subject-matter; it is at least an equal test of the thought-power or intelligence of pupils, since this result, in both methods, is dependent upon the nature of the tests; and, lastly, the certainty of the coming written test affords a healthy stimulus to pupils, increasing their attention to instruction, and their efforts to master the subject taught. (Ross & Stanley, 1954, p. 29)

By the end of the 19th Century, certification by written examination had fairly well caught on not only in England and on the Continent but also in the United States. And, as the legacy which the early Chinese provided for, now one could look toward "uniformity of testing situations" and "objectivity of appraisals" (DuBois, 1970, p. 10).

**Washburne's Winnetka Plan**

Carleton Washburne's so-called Winnetka Plan was an outgrowth of the educational measurements movement. And, as Hunt (1975) noted, "the tests used half a century ago in schools resemble more the new criterion-referenced tests than they resemble the norm-referenced ones" (p. 343). Similarly, the instructional practices advocated by Washburne and other contemporaries, such as Henry C. Morrison, resemble very much the mastery learning models advocated today by Carroll (1963), Bloom (1968), Mayo (1970), and others. Fortunately, Washburne's writings (e.g.,
Washburne, 1922, 1932; Washburne, Vogel, & Gray, 1926; Washburne & Marland, 1963) give us a fairly comprehensive picture of what the Winnetka Plan was like:

The first of these steps [of the Winnetka Plan] is the establishment of definite goals or subject-matter units. This is a natural outgrowth of the educational measurements movement. The second step is also a development of the same movement—the preparation of tests which will completely cover each subject-matter unit and diagnose the difficulties of each individual child. The third step is the preparation of self-corrective practice materials which will at once prepare for those tests and enable a child to make up the deficiencies shown by the tests. When one has taken these three steps there is no difficulty whatever in placing an entire school system on an individual basis. Achievement units become the constant factor almost automatically. For after all, the time unit is the arbitrary one, the achievement unit the natural one. When achievement replaces time as the constant factor in our school systems, we can promote children individually by subjects and fit our public schools to the needs of the individual child. (Washburne, 1922, p. 206)

Thus, Washburne's Winnetka Plan required (a) goals, (b) diagnostic tests, and (c) self-instructional, self-corrective materials for remediation; or in today's terms, a paradigm of Learning Objectives—Diagnosis—Prescription—Remediation. Washburne's plan called for individualized instruction and, like Carroll's (1963) Mastery Learning Model, the varying of time needed to achieve mastery of each instructional unit.

Washburne was director of tests and measurements at San Francisco State Normal School when that faculty began development of the tests that would be used in the Winnetka Plan. In 1919, as Washburne became superintendent of the Winnetka (Illinois) Public Schools (a post he held until 1943), work on the test development was continued by the Winnetka school faculty. Each test developed for the Winnetka Plan had to conform to four
criteria:

1. It must cover completely every detail of the unit tested—e.g., an addition test must include every combination, a short division test every type of difficulty, etc.;
2. the test must diagnose readily the exact weakness of each pupil insofar as the weakness is a lack of knowledge or ability;
3. the test must be strictly objective and easily corrected;
4. the answers must be so keyed that the pupil can readily turn to corresponding practice material and make up his deficiencies. (Washburne, 1922, p. 200)

Washburne lamented that "educational measurements as now known to most of us, however, are not yet sufficiently developed to enable us to make the complete change in our schools which is implied by variable time units and constant achievement units." But Washburne felt the movement toward this change "unmistakable" (p. 195).

Interestingly, Washburne and his associates apparently did not themselves routinely refer to the Winnetka techniques as the "Winnetka Plan":

To us . . . there is no such thing as the "Winnetka Plan." A "Winnetka Plan" would imply a certain fixity of organization, a setting up of a particular scheme as a model to be followed by others. Such organization would be contrary to the policy and spirit of the Winnetka Public Schools. For these schools are organized as a laboratory for scientific research and practical experimentation, and they are continually modifying their procedures in terms of their findings. (Washburne, 1932, pp. v-vi)

Morrison's University of Chicago Laboratory School Approach

As professor of education at the University of Chicago, Henry C. Morrison authored a book, first published in 1926, on his University of Chicago Laboratory School strategy (Morrison, 1931). His approach was not altogether unlike that of Carleton Washburne and his Winnetka Plan.
Like the Winnetka Plan, the Laboratory School strategy involved mastery learning. For Morrison, mastery was all-or-none:

When a student has fully acquired a piece of learning, he has mastered it. Half-learning, or learning rather well, or being on the way to learning are none of them mastery. Mastery implies completeness; the thing is done; the student has arrived, as far as that particular learning is concerned. There is no question of how well the student has mastered it; he has either mastered it or he has not mastered it. It is absurd to speak of degrees of mastery as to speak of degrees in the attainment of the second floor of a building or of degrees in being on the other side of the stream, or of degrees of completeness of any sort whatever. The traveler may indeed be part-way across the stream, he may be almost across, but he is not across until he gets there. Once across, he may continue his journey indefinitely, but he cannot continue his journey from midstream. (p. 36)

Note that Morrison did not rule out degrees of learning, but he refused, rather strongly, to call these mastery learning: "... in the unit learning itself there are no degrees; the pupil either has it or he has it not" (p. 36).

Morrison's (1931) paradigm of instructional attack included: (a) identification of the units of learning and (b) application of his "mastery formula": "pre-test, teach, test the result, adapt procedure, teach and test again to the point of actual learning" (p. 81). Morrison showed the analogy of this procedure to the physician and the agriculturist. The physician, for example, in attempting to cure his patient:

... first makes his diagnosis, then formulates and applies a treatment, then tests the results of his treatment, modifies treatment in accordance with his test results, and so on to success or failure. Even if he fails, the physician is eager to know why he failed. He does not merely dismiss the case with the verdict, "Failed to recover," or, in performance technology, "Failed to pass." (p. 81)
Again, like Washburne's program, Morrison's included learning objectives, diagnosis, prescription, remediation, and mastery. And, like the Winnetka Plan, the Laboratory School strategy allowed for variable time in order that the pupils achieve the particular learning unit.

**Other American Roots**

Criterion-referenced measurement and mastery learning have other American roots dating back to the early 20th Century. For example, Ballow (1916, p. 62) reported on the work of the Department of Educational Investigation and Measurement of the Boston Public Schools which required Boston teachers to compile a list of words that all students be able to spell by eighth grade. Also, English requirements were designated in behavioral terms. However, students' performance was viewed against city-wide standards, which is, or course, norm-referenced.

E. L. Thorndike (1918, p. 18) presented a distinction between two kinds of educational measurements: (a) one, relating to the psychologists' method of average error, which measures how well a student performs a task and (b) one, relating to the method of right and wrong cases, which measures how hard a task a student can perform at a specified level of success. Airasian and Madaus (1972, p. 2) called the former an example of norm-referenced measurement, the latter an example of criterion-referenced measurement. Earlier, in 1913, E. L. Thorndike discussed relative grading and his preference for absolute scores as an indication of an individual's proficiency.
In 1922, Helen Parkhurst authored a book on education on the Dalton Laboratory Plan (first introduced by Rosa Bassett in London) which provided for contracts between teachers and students for specified assignments and for varying the amount of time necessary to complete the assignments.

Test publishers, too, began realizing the necessity for considering the needs of the individual child. According to the 1932 Supervisor's Manual of the Metropolitan Achievement Test:

We are in a period of educational history when emphasis on the individual child predominates. This emphasis will be but a meaningless slogan unless we know the strength and weakness of the child and plan intelligently for his particular needs as far as class and other limitations will permit. (Cited in Fitzgibbon, 1975, pp. 4-5)

Block (1971, p. 4) noted that there was a void in the idea of mastery learning after the Washburne and Morrison plans in the 1930s until the late 1950s and early 1960s when the idea reappeared as a corollary of programmed instruction. But Knapp (1974) did not view the intervening period as a total void. In fact, Knapp traced roots of the criterion-referenced movement back almost exclusively to the period after the Washburne and Morrison writings:

The criterion-referenced testing movement can, perhaps, be traced back as far as the development of learning theories (Skinner, 1938; Hull, 1945). These theories pointed to the importance of individual differences in learning and encouraged the detailed analysis of both simple and complex tasks. The military put these theories into action during World War II by individualizing instruction through the use of programmed learning experiences. (p. 1)

Recent impetus for the criterion-referenced testing movement, according to Knapp (1974), has come from the development of instructional
If criterion-referenced measurement has such a rich history, why has the emphasis on this topic among educators and among educational test publishers been comparatively recent? It is easier to report on the reasons for the new emphasis in the field rather than on the previous lack of emphasis. But Block (1971) believes that the idea of mastery learning temporarily died after the 1930s due mainly to the "lack of technology required to sustain a successful strategy" (p. 4). Ralph Tyler (1976) has a more intriguing theory, however:

The testing movement had other effects that now appear to be negative. The promising work of Morrison and his colleagues on mastery learning was not followed up until forty-eight years later, because many educational leaders accepted the idea that a normal curve of educational achievement was to be expected and that the mastery of something by all the students must be abnormal and, therefore, wrong. (p. 30)

The irony of Tyler's view is even more striking when one recalls Washburne's giving partial credit for the impetus of the Winnetka Plan to the educational measurements movement. Thus, it would appear that not only did the early advances of educational measurement foster the educational innovations of Washburne, Morrison, and others, but the same advances apparently prevented these innovations from reaching the fruition that they finally are reaching today.
CHAPTER III

TECHNICAL ISSUES

Since the new emphasis on criterion-referenced measures, much discussion has revolved around whether these measures fit traditional test theory models. While some have argued for the application of traditional methods of test construction, including item tryouts, item analysis, reliability and validity studies, etc., others have stated that none of these is possible and that face validity and content validity are the only necessary requirements; yet still others have developed their own theories, their own models, and their own coefficients.

If there is anything that can be agreed upon in the field of criterion-referenced measurement—and this field includes diagnostic and mastery learning and testing, it is that there are many unresolved issues. In fact, there seem to be many more papers dealing with questions than there are dealing with answers. And, worse, there are a number of papers by investigators who go blithely about their business not realizing that they are in an area with many unresolved problems.

Problems and Questions

Boehm (1973), in noting that "criterion-referenced testing is not a panacea for all the problems encountered in assessment," listed what she
considered to be the problems:

1. Who determines the objectives?
2. Who sets the behavioral criterion levels?
3. Do test items accurately reflect the behavioral criteria?
4. What constitutes a sufficient sample of criterion levels?
5. Do the test scores obtained describe an individual’s response patterns?

In addition, there is the lack of accepted theory and procedures for determining test reliability and validity. . . . (p. 120)

R. L. Baker (1974) believes that concern with certain psychometric dogma is misplaced and nonproductive. He asked:

1. Is the criterion-referenced test just a special instance of the norm-referenced test?
2. How can the reliability of criterion-referenced tests be assessed? (p. 37)

Baker believes that the classical tools of validity, reliability, item analysis, norming, etc. are necessary but that "more sophisticated tools are clearly needed" (p. 45).

Harsh (1974) saw technical problems in both norm-referenced and criterion-referenced tests:

It is recognized that there are many technical problems involved in using either norm-referenced or criterion-referenced tests for making conclusions about the true growth of student populations. . . . Suffice it to say, the reliability and validity of the measures are troublesome problems that plague those interested in very precise and parsimonious conclusions concerning short-term, annual, or longitudinal growth in academic achievement. (p. 10)

One of the problems in criterion-referenced measurement is the lack of variability about the test scores. If a student does not know the subject matter, he or she should theoretically not get the test item(s) correct; if the student knows the subject matter, he or she should, theoretically, get the
item (s) correct. The concept of variability will be dealt with in greater
detail later in this chapter. For now, it is interesting to point out that Jack-
son (1971) noted Popham's and Husek's (1969) argument that conventional
procedures for item analysis and assessment of reliability and validity are
not applicable to criterion-referenced tests because test scores on a crite-
ron-referenced measure may have no variability in the population of inte-
rest. This lack of variance does not, however, imply that the test is not
good, useful, reliable, or valid (Jackson, 1971, p. 10).

Kifer and Bramble (1974) observed that:

... with increasing frequency the psychometric properties and prob­
lems of criterion-referenced tests are appearing in the research litera­
ture. Much of the discussion of these tests focuses on the extent to which
classical test theory is an appropriate perspective from which to view
criterion-referenced measurements.

They then posed three problems of criterion-referenced tests:

1. Assuming that a standard or criterion has been chosen, how general­
izable is it?
2. What is the relationship between the test items and test scores and
those standards?
3. Given that the scores are compared to a standard, with how much
precision can one state whether a particular score represents attain­
ment or above the standard?

Klein and Kosecoff (1973) felt that criterion-referenced test devel­
opers must clarify the nature and purpose of a criterion-referenced test by
answering a number of questions, including the following:

1. For what decision areas and purposes is the CRT [criterion-refer­
enced test] most applicable?
2. What areas and objectives does the CRT cover and how were these
objectives derived and organized?
3. How broadly or narrowly are the objectives defined?
4. How were the test items or tasks chosen to measure the objectives defined and developed?
5. How dependent are the items on particular instructional materials or programs? And what is their applicability to different kinds of students?
6. What methods were used to improve the items on the CRT and why were they chosen relative to the purpose of the instrument?
7. How was the validity of the CRT established?
8. What kinds of scores should be reported for a CRT and what is the justification for these scores, especially those involving "mastery"?
9. How was the test finally put together, what compromises had to be made, and how were they resolved?
10. In what ways will packaging of the CRT facilitate its use? (p. 15)

Nitko (1974) saw procedures needed for the following unresolved problems in criterion-referenced measurement:

1. Defining the behaviors to be taught and tested for in the instructional situation
2. Task analysis as it relates to school-like behaviors
3. Relationship between what is tested and the ultimate objectives of the individual and society
4. The relationship between the behavioral domain and the domain of tasks serving as the potential item domain
5. Specification of the domain of tasks in terms of their stimulus and response characteristics
6. The ordering of the domain of behaviors in terms of their psychological structure
7. Data related to the generalizability of samples of behavior to the behavioral domain
8. Construct validation of proposed orderings of the behavioral domain
9. The development of an item-writing theory and an item-response theory
10. Development of procedures for determining mastery of identified behavior (p. 78)

Harris (1974a) saw these problems:

1. What objectives are to be reached?
2. How are the objectives to be written or formed in order to provide bases for instructional development, and/or bases for measurement procedures?
3. How are the measurement procedures to be developed?
4. How are the measurement procedures to be used? (p. 84)
Popham (1974) posed these two questions:

1. Which objectives should be selected for inclusion in the tests?
2. How should test items be constructed so that they will be homogeneous representatives of the test-item domain circumscribed by an objective? (p. 14)

Then, Popham (1974) postulated the following five unresolved problems:

1. What techniques can be devised which will permit objectives-based test developers to improve their instruments on the basis of empirical tryouts in the same ways that conventional test developers have been doing it for years (e.g., total test reliability, item reliability, item homogeneity, objective-item congruence)?
2. How can a replicable set of guidelines be produced which will allow one to economically yet definitively constrain item-writers who will produce objectives-based tests?
3. Are there technical rules which can be produced to aid reviewers in judging the congruence between test items and the objectives on which they are based?
4. Can a technology be devised to assist objectives-based test designers to delineate satisfactory criteria so that items calling for constructed learner responses can be employed with the expectation that the resulting responses can be reliably scored?
5. Was our decision defensible to devise tests which assess only certain objectives (for example, X₄) versus sampling from many objectives (X₁, X₂, X₃, X₄, etc.)? (p. 25)

Skager (1974) saw six unresolved questions in regard to developing criterion-referenced tests in such a way that they address various information needs in education:

1. **Independence:** Given the rationally derived structure of a content domain, a set of performance objectives devolving from that structure, and pools of assessment materials written to measure each objective, is there any need to verify empirically whether or not the performances specified by the objectives are sufficiently independent from one another to provide non-redundant information?
2. **Validity:** How does one establish the fact that the items in the pool measuring any objective are valid in the sense of being (a) congruent with the objective, e.g., actually measuring the performance described in the objective, and (b) comprehensive in the sense of providing adequate coverage of the domain specified by the objective?
3. **Identifying "Bad" Items:** How does one identify poorly written items
by means of item analysis procedures when the frequency of correct response may be extremely high or low, accurately reflecting the achievement status of a particular group of learners?

4. **Information on Items in Bank:** Assuming that the items in a bank have met necessary tests of quality, what sort of information might be stored on each that would aid in constructing tests and interpreting the scores which would eventually result?

5. **Sequencing Objectives:** When the collection of objectives represents terminal points in instruction is it necessary and appropriate to find some "ideal" sequence by which instruction might proceed?

6. **Defining Mastery:** How many items does one include on criterion-referenced tests when the purpose is to determine whether learners have achieved mastery of an objective (or objectives), taking into account (a) the generality of the item pool in terms of the variety of performances defined by the objective, (b) whether the response called for is to produce the right answer or select the right answer, and (c) whether the resulting information will refer to individual learners or groups of learners? (pp. 51-55)

Wilson (1974), in a discussion on National Assessment, saw two problems:

1. What constitutes a definition of a domain of reference of a universe of behaviors?
2. When can we be sure that a complete definition is achieved? (p. 28)

There are many more questions and problems, and proposed solutions, posed by the above investigators and others. And, of course, many of the above questions are merely restatements of the same concerns. Sometimes, the restatements are not immediately obvious because of the different frames of reference of different investigators (e.g., criterion-referenced versus domain-referenced). When one sorts through the many questions, problems, issues, etc. in the field of criterion-referenced measurement, the more common ones appear to be the following; and these will be elaborated on in more detail in the remainder of this chapter:
1. Behavioral Objectives
   Writing Objectives
   Selecting Objectives

2. Test Construction
   Item Writing
   Number of Items Per Objective
   Item Analysis

3. Reliability

4. Validity

5. Mastery Criterion

6. Norm-Referenced Criterion-Referenced Measures

The above issues are not meant to be exhaustive; but elaboration on
and resolution of the above issues will go a long way toward moving the
field of criterion-referenced measurement toward a more mature stage of
development.

**Behavioral Objectives**

Before a teacher can teach his or her students something or before
the teacher can test something, clearly-defined goals must be present. The
Gileadites, authors of one of the earliest criterion-referenced tests, had an
important, clearly-defined goal—to identify friendly Gileadites and distinc-
tuishes them from enemy Ephraimites. The ancient Chinese, who, according
to DuBois (1970, p. iv), invented the psychological test, had a goal—to
identify those fit for service in governmental posts. Morrison's University
of Chicago Laboratory School strategy and Washburne's Winnetka Plan both indicated the need for setting instructional goals as the first step toward implementing their programs. Generally, it can be stated that individuals do have goals or objectives which can influence their behavior. In instruction and testing, one major goal is to determine more specific goals so that the instructional process will be a meaningful one and not a random one.

Objectives exist on various levels of specificity. It is not so important to determine how many levels of objectives there exist in nature as it is to come up with a set of objectives that can be dealt with in a day-to-day operation. But to give some insight into the possibilities of levels of objectives, some examples will be given.

Popham (1974) reported on a four-level hierarchy of instructional objectives used at the Instructional Objectives Exchange (IOX) at UCLA:

1. **Major Categories**—important and comprehensive skills
2. **Content General Objectives**—intermediate skills
3. **Objectives**—precise statement of a skill
4. **Amplified Objectives**—expanded objectives containing sufficient detail regarding the nature of measurement procedures to facilitate item development (p. 14)

Wilson (1974) used a three-level hierarchy at the National Assessment of Educational Progress (NAEP):

1. **Overall Objectives**—educational goals
2. **Major Objectives**—specific content areas and behaviors
3. **Sub-Objectives**—precise performance criteria (pp. 29-30)
In general, one can characterize a hierarchy of objectives via the following example:

1. The teacher wants to diagnose skill weaknesses in the curriculum.
2. The teacher wants to diagnose skill weaknesses in reading.
3. The teacher wants to diagnose skill weaknesses in phonics.
4. The teacher wants to diagnose skill weaknesses in initial consonants.
5. The teacher wants to diagnose skill weaknesses of initial consonant b.
6. The teacher wants to diagnose whether his or her students can identify the letter/sound relationship of the initial consonant b.

In very simplistic terms, the above is all the hierarchy-of-objectives issue involves. The issue is not so much a controversy as it is a fact of life. In instruction and in measurement, one would ordinarily start with the general and proceed to the specific. Similarly, the distinction often made between ultimate versus immediate (or proximate) objectives (e.g., Krathwohl & Payne, 1971, pp. 18-20) merely provides labels to existing concepts.

Writing Objectives

What is probably more of an issue than labeling types of objectives is how to word objectives and how precise the wording should be. This issue is more of a problem than a controversy. This writer has had the opportunity to view many sets of behavioral objectives from various curriculum guides, district and state programs, testing programs, etc., with a goal toward "correlating" these sets of objectives to each other. The
varied nature of the wording and the preciseness of each set of wording makes this correlation a very difficult task. It would be easy to say, for example, that District A's criterion-referenced test and District B's criterion-referenced test both measure contractions if each test has items to measure contractions. But what if District A's objective states, "Given the two words that make up a contraction, the learner will be able to supply the correct contraction" and if District B's objective states, "Given a contraction, the learner will be able to supply the two words that make up the contraction"? Does the investigator call these two objectives a match? If the investigator says "yes," then he or she has correlated two objectives that have the processes reversed; in this simple example, one can see that these two objectives are not perfectly matched. But, if the investigator says "no," and if these are the only objectives relating to contractions in the two tests, then the implication would be that the two tests do not cover the same content, when in effect they do. The above is a dilemma which really does occur. Matters would be simplified if curriculum guides would be more generally written, leaving the precise statements to the evaluators to define a testing process as opposed to an instructional process.

Alkin (1974), referencing Mager (1962) and Popham (1965), stated that two elements that may be covered in an instructional objective are:

1. A definition of conditions under which measurement of the objective takes place (e.g., open vs. closed book; in front of a student audience vs. into a tape recorder)
2. The specification of standards of performance to be reached in order for the objective to be achieved (e.g., 80% correct; in less than two minutes) (Alkin, 1974, p. 8)
Skager (1974) stated two essential characteristics of performance objectives:

1. An action statement describes an observable behavioral output.
2. There is a description of the conditions and materials with which the examinee is to perform the action. (p. 47)

Most of the discussions on writing objectives are influenced by or are similar to Robert Mager's (1962) classic exposition on the subject, Preparing Instructional Objectives. Mager presented essentially three rules-of-thumb which are worth noting here:

1. Identify the terminal behavior: What will the learner be doing?
2. Define the desired behavior further: What conditions will you impose?
3. Specify the criteria of acceptable performance: How will you recognize success?

The following is an adaptation of the major points covered in Mager's (1962) book:

I. Definition of Terms

A. Behavior--refers to any visible activity displayed by a learner (student)

B. Terminal behavior--refers to the behavior the learner should be able to demonstrate at the time the instructor's influence over him or her ends

C. Criterion--is a standard or test by which terminal behavior is evaluated
II. General Procedure of Writing Objectives

A. First, identify the terminal behaviors by name; specify the kind of behavior that will be accepted as evidence that the learner has achieved the objective. (What will the learner be doing?)

B. Second, try to define the desired behavior further by describing the important conditions under which the behavior will be expected to occur. (What conditions will you impose?)

C. Third, specify the criteria of acceptable performance by describing how well the learner must perform to be considered acceptable. (How will you recognize success?)

III. Identifying the Terminal Behavior

What is the learner doing when he or she is demonstrating that he or she has achieved the objective. For example:

To be able to solve quadratic equations
To be able to repair a radio

IV. Further Defining the Terminal Behavior

State the conditions that will be imposed upon the learner when he or she is demonstrating his or her mastery of the objective. For example:

Given a matrix of intercorrelations . . .
Given a list of . . .
Without the aid of references . . .

V. Stating the Criterion

Tell the learner how well you want him or her to be able to do the
objective. For example:

within a time limit
minimum number of correct responses
accuracy

In writing behavioral objectives, it is important to use action verbs which are not open to many interpretations. Mager listed the following as words open to many interpretations: to know, to understand, to really understand, to appreciate, to fully appreciate, to grasp the significance of, to enjoy, to believe, to have faith in. The following are words open to fewer interpretations (and hence are more desirable to use in formulating behavioral objectives): to write, to recite, to identify, to differentiate, to solve, to construct, to list, to compare, to contrast.

Finally, how can one test whether a written objective clearly defines a desired outcome?—Can another competent person select successful learners in terms of the objective so that you, the objective-writer, agree with the selections. If the answer is yes, the test is fulfilled (Mager, 1962).

Selecting Objectives

Selecting behavioral objectives for use in a criterion-referenced test is a difficult, challenging process. As R. L. Baker (1974) pointed out:

The "how-to" information for stating well-formed instructional objectives has been available for some time. . . . However, the time-consuming and thought-challenging task of what outcomes to prepare remains to be done. But this is a matter of doing the job, rather than of not knowing how. (p. 38)

Proper selection of objectives is crucial because the objectives determine criterion-referenced test content, item selection, and score
interpretation (Skager, 1974, p. 47). Skager argued that test constructors should keep in mind the various assessment and information needs while building a test:

1. **Planning the curriculum**: What is the content of the tests which will later be used to assess the effectiveness of instruction?
2. **Classroom Management**: What is the present learning status of the pupils in terms of the objectives and prerequisites of the instruction?
3. **Evaluating Instruction**: What is the terminal learning status of students who have been exposed to the program or one or more of its sub-units?
4. **Accountability**: What is the terminal learning status of students who have been instructed by particular teachers or in particular schools?
5. **Allocating Resources**: Where are the deficiencies in the achievement of students so severe as to require the allocation of additional efforts and funds?
6. **Prediction**: What will be the future achievement of individuals in particular educational or employment situations? (p. 48)

There seems to be some advantage in maintaining control over the selection and writing of objectives, rather than farming out these tasks to outside agencies. Wilson (1974) reported that, in the early years of National Assessment, subcontractors (for example, American Institutes for Research, Educational Testing Service, Science Research Associates) developed objectives and wrote exercises to measure these objectives. This procedure "not only produced objectives of uneven quality but was also liable to produce only those objectives that were most easily measured while neglecting those that are difficult to measure but still important to the education community." Later, an Exercise Development Department of NAEP was charged with producing objectives and test items (Wilson, 1974, p. 31).
Sullivan (1973) presented eleven considerations in selecting and using instructional objectives:

1. Is it really important for the student to possess the skill stated in the objective?
2. Does the student already possess the skill stated in the objective?
3. Is the skill or attitude teachable?
4. Should I set performance standards?
5. Is the student given the information he needs in order to attain the objective?
6. Do the instructional activities provide direct practice on the objective?
7. Does the learner possess the prerequisite skills necessary to attain the objective?
8. Should there be alternative instructional methods for the objective?
9. Are the instructional activities potentially appealing?
10. Does the planned assessment measure the skill stated in the objective?
11. Does the importance of the objective justify the estimated time and expense? (p. 2)

Popham (1974) presented sets of criteria for selection of major categories of objectives and of specific objectives. For major categories, Popham's suggested criteria included:

1. Importance. What topics, what skills, etc., will be viewed by educators as most important for that subject?
2. Economy of Production. What topics, content, skills, etc., can we translate into tests rather readily?
3. Practical Scorability. Which major categories and content general objectives associated with them are apt to yield specific objectives and resulting test items which will be readily scorable? (p. 15)

For specific objectives, Popham's suggested criteria included:

1. Widely accepted. The objective selected should be the most widely accepted as important by those in the field.
2. Transferability within Domain. The form of learner behavior selected should be the most generalizable of those represented in the content general domain, i.e., a learner mastering the designated behavior requirements would likely be able to transfer that mastery
to most, if not all, of the other eligible behavioral requirements in the content general domain.

3. **Terminality**. . . . The chosen specific objective should represent the most terminal learner behavior.

4. **Transferability Outside the Domain**.

5. **Ease of Scorability**.

6. **Amenability to Instruction**. (p. 17)

The actual method used to select objectives could vary considerably.

**Gronlund (1973)** recommended keeping in mind two major considerations for selecting objectives:

1. **What should be mastered in a particular learning situation?**
2. **What can be mastered in a particular learning situation?** (p. 8)

**Guzaitis (1973)** provided practical advice on how a teacher can prepare his or her own criterion-referenced tests:

By analyzing the basal text and/or curriculum guide for each subject, the teacher can formulate an outline of those skills that he wants his students to have mastered by the end of the unit, course, or year. This outline should also show where these skills are presented in the textbook (by page number), skill kit (by card number), tape program (tape and side number or name), and so forth. . . . After compiling this outline of skills and prescriptions, the teacher should state each skill in terms that will clearly reveal when the student has mastered it. This will typically result in statements that begin with a "given," proceed to "be able to," and terminate with an "outcome" that not only the teachers but others can recognize as fulfilling the requirements set forth.

**Wilson (1974, p. 31)**, in discussing National Assessment, recommended searching recent literature to identify new trends in the subject area and examining existing sets of written objectives, e.g., the IOX objectives and others. Wilson also noted that NAEP had chosen to take a judgmental rather than a statistical approach to defining the set of objectives to be measured (p. 29).
Judgment would also be involved if a curriculum coordinator or an administrator or a teacher selected the objectives. Judgment is necessary even if the textbooks have a scope and sequence chart, though even more judgment would be required if objectives were selected and written based on analyses of textbooks' tables of contents or on page-by-page analyses of skills taught. One way of sharing the burden of these judgments would be to submit the objectives selected to a panel of experts—subject-matter and test specialists. Popham (1974, p. 23) noted that such was done for the IOX tests and Wilson (1974, p. 31) noted that such was done for NAEP tests. Test publishers generally do likewise in their development of assessment measures.

**Test Construction**


**Item Writing**

After defining the objectives to be measured in an achievement or a mastery test, the next step is to "write items to sample content and behavior domains of the objective" (Mayo, 1970, p. 2). Guzaitis (1973) added: "If the objectives are well conceived, the items will write themselves; if not, they will be difficult to construct. Thus, you have a cross-check on
the clarity of the objectives.

For generating items, Popham (1974, p. 18) suggested the use of an amplified objective which delimits the stimulus elements and describes the learner response options, the goal being to help the item-writer produce homogeneous items. Popham added that the amplified objective must "limit meaningfully the set of eligible test items without, at the same time, trivializing the set of items" (p. 19). Davis and Diamond (1974) agreed with Popham in terms of homogeneous items; they distinguished a diagnostic test as one that must be composed of a series of homogeneous items from a survey test as one which would be composed of heterogeneous items (pp. 120, 122). Macready and Merwin (1973) added:

For an item to have the most desired relationship with other items in a set, two conditions must be present. First, all items within the item form need to be of equal difficulty. Second, there should be homogeneity between these items. (p. 353)

(The concept of equal difficulty will be returned to later in this chapter.)

Klein and Kosecoff (1973) provided advice for putting together criterion-referenced tests:

1. Combine objectives that are considered highly related to one another into a single measure.
2. Select a group of objectives from the total pool of objectives based on a set of appropriate criteria.
3. Limit the scope of each objective so as to reduce the potential number of items and/or tasks that might be needed to measure it.

(p. 9)

How many alternatives should a multiple-choice item have? A typical standardized achievement test will have four, maybe five, choices for each item. Costin (1972) compared three-choice versus four-choice multiple-
choice test items with implications for reliability and validity of achievement tests. Using an introductory psychology course (N = 1,566), Costin selected 100 four-choice test items, randomly selected 50 of them, and randomly dropped one of the distractors of these 50 items. Costin calculated Kuder-Richardson formula 20 estimates of homogeneity for each set of 50 items, the means of the point biserial correlations between test items and total test score, the mean number of items answered correctly, the standard deviation, the median of the scores, and the standard error of measurement for each set of 50 items. Results showed virtually no difference between the various measures (Costin, 1972, p. 1037). Costin recommended that because of the virtually identical results when using three-choice and four-choice multiple-choice items classroom achievement tests use three-choice items--for increase in efficiency, with less testing time required. Costin felt that more studies were needed to empirically determine the relationship between number of item alternatives and reliability and validity of these tests (p. 1038). (Note: Costin (1972, p. 1036) reported that Horst (1966) had suggested that the point biserial correlation coefficient might be considered a measure of homogeneity (rather than discrimination power) because the higher the point biserial, the greater the degree of homogeneity with the rest of the items.)

A more recent paper by Grier (1975), using a modified version of the Kuder-Richardson formula 21, tended to support the theoretical advantages of three-choice multiple-choice tests:
But this is true only if the number of test items is increased to compensate for the smaller number of alternatives per item. For example, suppose that in a test hour it is reasonable for students to complete 60 four-alternative items. A test that is more reliable, more powerful, more discriminative, and more informative can be achieved by switching to 80 three-alternative items. This argument assumes that the additional test items are available and that the test can be finished in about the same time. (p. 112)

Though there is no question that it is easier to construct two distractors versus three distractors (often test developers seem to strain to come up with that last distractor), the theoretical chance level of the test ought to be taken into consideration. As one of this writer's former professors used to define it, the theoretical chance level is the "score that would be achieved by a blind ape with a pencil"; that is, random marking. Therefore, the blind ape with a pencil would be expected to score only 25% on a test with four choices, but a more respectable 33% on a test with three choices. It would take a great deal of persuasion, despite some studies to the contrary, to convince this writer to abandon his preference for four choices in a multiple-choice test.

What about rules for item generation in criterion-referenced testing? Wilson (1974) believes that "it would be desirable to identify a generally acceptable method for item construction" in criterion-referenced testing.

The method should provide:

1. A systematic sampling of a previously defined universe of behaviors
2. A set of rules which, if followed by more than one person or group of item writers with equivalent knowledge, would produce equivalent tests (pp. 32-33)

In regard to point one, Wilson noted that the universe of behaviors has not
been well defined; in regard to point two, Wilson noted that this set of rules is more useful in narrowly specialized areas than in other more complex areas (p. 33).

Wilson (1974) went on to describe the method of test development used by NAEP. NAEP used subject-matter experts experienced with students at the four age levels to weight the major objectives and each sub-objective for relative importance; the weights were translated to amount of exercise material developed for each sub-objective (pp. 33-34). A panel of subject-matter experts selected appropriate item prototypes, specified time ranges, and produced "exemplary" items for the subject area. Contractors then developed exercises and provided rationale for tying in the item to the sub-objective and to other items in the test. The items were reviewed by the NAEP staff, subject-matter experts (scholars and educators), and qualified lay persons. The items were then tried out in extreme inner city, extreme rural, and affluent suburban areas (pp. 34-35). In terms of item prototypes, NAEP provided a tree structure resulting in 80 \((2 \times 4 \times 2 \times 5)\) possible item prototypes:

1. Administrative Mode (2)
   --Individual or Group
2. Stimulus Mode (4)
   --Audio; Visual; Other Senses (Tactual, Olfactory, Etc.);
   Combination of Three
3. Response Mode (2)
   --Objective (Multiple-Choice) or Free Response
4. Response Category (5)
   --Written; Verbal; Role Playing; Group Interaction; Other Physical Action (Wilson, 1974, p. 34)
At this point, it might be useful to provide a brief discussion of Hively's work with item forms. Nitko (1974) defined Hively's item forms analysis as:

... the process whereby behavioral statements are analyzed in order to derive classes of items which elicit the various aspects of the behavior class. As a result of this analysis, one or more item forms are derived for each behavior class. (p. 65)

Hively prefers to make a distinction between domain-referenced testing and norm-referenced testing; he does not like the term "criterion-referenced testing" because "the term 'criterion' lends itself to misinterpretation. It carries surplus associations to mastery learning that are best avoided by using the more general term 'domain' instead" (Hively, 1974, p. 5). Now, let's present three of Hively's (1974) definitions that are relevant to item generation:

DEFINITION: An "item" is a set of instructions telling how to evoke, detect and score a specific bit of human performance. It must include directions for (1) presenting the stimuli, (2) recording the response and (3) deciding whether or not the response is appropriate. (p. 6)

DEFINITION: A "domain" may consist of any clearly specified set of items. (p. 8)

DEFINITION: A list of rules for generating a set of related items is called an "item form." (p. 8)

An item form consists of a specification of the invariant part of the class of items together with (a) an indication of which parts of the items are variable, (b) a specification of elements which can be used in the variable parts of the items, and (c) a specification of the rules by which one selects an element from the set of variable elements to derive a particular item. . . . The variant part of the item is called a shell; the sets of elements which can be used in the variable parts are called replacement sets; and the rules by which one samples from the replacement sets are called the replacement structure. . . . (Nitko, 1974, p. 65)
The goal of item forms is to enable item writers to construct items based on explicit rules, rules which should leave no doubt as to the composition of the item itself. In practice, such a technique would tend to work better for a simple mathematics objective than, say, a complex reading objective. The task can be a difficult one; and Popham (1974) stated that Hively's system "has, for some, proved too sophisticated for sustained use" (p. 18).

Several other investigators have presented principles of criterion-referenced test item generation:

1. The keyed response must be an adequate correct response—not merely the best of the responses included.
2. All distractors must be clearly incorrect or (in best-answer items) generally accepted by informed authorities in the field as less adequate answers from the keyed response.
3. Distractors should be as attractive as the psychological context of the item permits and should be as nearly equally attractive to examinees in the target population as possible; that is, each distractor should attract as nearly as possible the same proportion of those examinees who cannot identify the correct answer.
4. Choices for an item should be logically coordinate and distractors should not overlap each other or be related in a way that allows one or more distractors to be eliminated by an examinee who is test-wise and can reason well but has no information or skill in the variable that the item is intended to measure. (Davis & Diamond, 1974, pp. 123-127)

According to Klein and Kosecoff (1973):

1. The plan used directly affects the utility, content validity, and score interpretation of criterion-referenced tests.
2. Relative difficulty of items with an objective affects score interpretation. Often a slight change in the item itself can modify its difficulty. And, "if only the most difficult items are used, then the phrase 'mastery of
the objective' has a very different meaning than if the items were sampled over the full range of difficulties."

3. The degree to which an item reflects a particular curriculum or set of materials and techniques affects generalizability of the scores—instructionally dependent or biased versus instructionally independent.

4. There is potential interaction between the test objective and how the item measures it (Klein & Kosecoff, 1973, pp. 5-6).

Number of Items Per Objective

Klein and Kosecoff (1973, p. 4) noted that "even a highly specified objective could have a potential item pool of well over several thousand items. . . ." And Raju (1975, p. 5) observed that the mathematics objective of "the learner will add two 2-digit numbers" has an item domain of 8,100 items. Thus, the problem results: How many items per objective should be tested? Factors which influence the number of items to construct for a given objective include "testing time available" and the "cost of making an interpretation error, such as saying that a student has achieved mastery when he has not" (Klein & Kosecoff, 1973, pp. 4-5).

Macready and Merwin (1973, p. 353) suggested that the ideal is for the student who gets one item in a population of items correct to also get other items in the population correct. Recall that these investigators also stated that all items within an item form should be of equal difficulty. Thus, as advocated by proponents of domain-referenced testing, where
items measuring an objective are to be randomly sampled from a domain of items, the item one sampled would indicate pass or fail of the objective. Macready and Merwin used the word "ideal," and the ideal is not likely to be met. While it might be possible to randomly select mathematics items from a finite domain of 8,100 items, for most objectives the domain of items is infinite and undefinable, except in trivial cases.

Fremer (1972) spoke about the problem of using a single item to measure an objective:

Whereas individual exercise reporting poses serious problems of interpretation because of the sampling error associated with the selection of only a single exercise, a set of homogeneous exercises tied to a single objective allows considerably greater accuracy. The size of the set will need to vary with the nature of the objective, however, to meet adequate measurement standards.

There are some theoretical models available to help determine test length. Millman (1972) offered an approach to determine test length of a criterion-referenced test via a decision rule which specifies mastery if the percent of items that a student answers correctly equals or exceeds the criterion level. But determining the criterion level—i.e., how many items correct determines mastery—is another issue altogether and is dealt with later in this chapter. Hence, Millman's approach, which uses binomial probability tables to obtain the probability that a student with a given true level of functioning would be incorrectly assumed to have achieved mastery, requires stating the unknown.

Novick and Lewis (1974) used application of Bayes' theorem to obtain the probability that a student has equaled or exceeded the criterion level,
given the student's test score. Novick and Lewis argued that their approach has advantages over Millman's (1972) because Millman's approach requires stating the true level of functioning (an unknown) while Novick and Lewis require stating the test score (an observable). But the problem with Novick's and Lewis's approach appears to be that the tables presented in their article require test lengths of at least seven items per objective to reach an acceptable criterion level—undesirable in terms of test time economy.

Klein and Kosecoff (1973) reported that a survey of current criterion-referenced tests "reveals that the usual practice is to use three to five items per objective. This practice appears to stem more from feasibility constraints than any sound foundation in psychometric theory or technology" (p. 5). Guzaitis (1973) provided additional considerations:

If the items are open-ended, fewer will be needed to measure attainment of each objective, although at least two should be used. If multiple-choice items are used, at least twice that number should be used for each objective to help control for guessing.

Novick and Lewis (1974) recommended that the number of items to test mastery of a behavioral objective be kept at a minimum but that the test be long enough to provide sufficient information about the degree of mastery of the behavioral objectives:

The minimum acceptable length depends on the manner in which test information is used to make decisions about individual students, the level of functioning required for defining mastery of an objective, the relative losses incurred in making false positive and false negative decisions, the background information available on the student and on the instructional process, and the premium on testing time within the instructional process. (p. 139)

Theoretical mathematical models aside, the best and most practical
approach appears, to this writer, to be the one adopted by Raju and his associates (SRA, 1975, p. 5). The approach adopted by Raju for Science Research Associates's *Mastery: An Evaluation Tool* (a component of the SRA Criterion-Referenced Measurement Program) involves three four-choice multiple-choice items per objective. Raju's rationale will be covered in a later section of this chapter, on the setting of the criterion for mastery.

**Item Analysis**

Item analysis techniques have several practical uses. They can be used to answer questions such as the difficulty of an item, whether the item discriminates between better and poorer students, and whether all the item alternatives are appropriately distracting.

Traditionally, in achievement test construction, the difficulty of an item is determined by the percent of students choosing the correct alternative. If, say, 20 students take a test and only 5 get item 13 correct, then item 13 has a difficulty, or $p$, of 5/20 or .25. Maximum variance of a test is achieved when $p$ values are at .50. In practice, a rule of thumb generally is to use items with $p$ between .25 and .75, with an average of .50.

Discrimination on an achievement test is determined by selecting a "better" and a "poorer" group of students. Often the top 25% and the bottom 25% of students (based on total test score) are chosen (although some research has indicated that the best breakpoint is the top 27% and the bottom 27%). Discrimination is computed by subtracting the percent of the lower group who choose the correct alternative from the percent of the upper
group who choose the correct alternative. A minimum discrimination of 15-20 is usually desirable.

A check on the item alternatives is to see that for each incorrect option, more lower than upper students select the option. Of course, more upper than lower students should choose the correct option!

Usually, test developers try out in a field test version 2.5 to 3 times as many items as they will need in the final version.

The item analysis procedure is not so simple for a criterion-referenced test, however. Raju (1974) observed that the major difficulty for criterion-referenced tests appears to be the variability of test scores:

In the Norm-Referenced Test (NRT) area, one attempts to build tests that can differentiate individuals within a group, and as such one capitalizes on individual differences or group variability. However, such differentiation is not only unessential but may indeed prove to be harmful in the CRT area. For example, let us say that an item dealing with the addition of two 2-digit numbers is administered to a group of students and that each student in the group has answered it correctly. This item has no value in a norm-referenced setting because everyone knows the answer to it and, hence, it cannot be used to differentiate students in the group. The same item, however, is useful in a criterion-referenced setting because it can help make inferences about each student's ability to add two 2-digit numbers. . . . Group variability, which is so essential to norm-referenced measurement, is at the core of the difficulties in CRTs. (pp. 1-2)

What this dilemma means is that the traditional type of item analysis may very well eliminate the best items for a criterion-referenced test. If 100% of a student population can correctly answer "2 + 5 = ?" then that item would have a p value of 1.00 and would be rejected by item analysis because the item is "too easy." But if the criterion-referenced test objective involved adding two 1-digit numbers, then the item is not too trivial; the data
merely indicate 100% mastery of that item. Similarly, an item that no one gets correct would ordinarily be rejected for a norm-referenced test; but if that item measures an important objective, the item is an important item and indicates nonmastery of that item.

As Popham and Husek (1969) pointed out:

With criterion-referenced tests, variability is irrelevant. The meaning of the score is not dependent on comparison with other scores; it flows directly from the connection between the items and the criterion. It is, of course, true that one almost always gets variant scores on any psychological test; but that variability is not a necessary condition for a good criterion-referenced test. (p. 3)

How do we resolve this dilemma? Perhaps by a judgmental approach.

E. L. Baker (1974) stated that "most decisions regarding content limits, criteria or distractor domains, formats, etc., are arbitrary, as are most curriculum decisions" (p. 16). What about experts, then, as judges of item difficulty? Item writers for National Assessment tests were asked to classify the items they wrote as very difficult \( p = .10 \), moderately difficult \( p = .50 \), or very easy \( p = .90 \). Subject-matter reviewers were then asked to concur or nonconcur with these estimates; in most cases they concurred. But when actual results for this test (Science) were reported, it was shown that the item writers and reviewers correctly classified the difficulties of the items only 68% of the time; "this is not outstanding success" (Womer, 1970, p. 7). Womer (1970) continued:

These results raise the question of whether it is possible for adults to do a really good job of estimating p-values of students. . . . The results . . . indicate that in the final analysis it is 9-year-olds who must tell us what 9-year-olds know, that it is 13-year-olds who must tell us what 13-year-olds know, and so on. (p. 7)
Though the above results are somewhat disturbing, a number of investigators seem to be advocating a judgmental approach toward criterion-referenced test development. While Womer's study suggested that it is difficult to judge item difficulty, the study does not suggest necessarily that it is difficult to judge item suitability. More empirical research is obviously necessary before drawing more definite conclusions. Meanwhile, some caution should be used in using a judgmental approach, which, by the way, may well be the best practical solution available at the present time.

Kriewall (1972) argued that the information needed from a criterion-referenced test is not the difficulty of the item but the proficiency score of the individual taking the test. For a norm-referenced test, item difficulty \( p \) is defined as "the expected relative score on an item by a population of examinees. . . . If an individual is selected at random from the population of examinees, then \( p \) is the probability such a person will respond correctly to the item." But a teacher is not selecting students at random from the population; the teacher is dealing with a particular group and must treat that group appropriate to its characteristics. Kriewall (1972) continued, "What the teacher needs to know at given points in time is the probability for success that a given pupil has with respect to a specified class of performance tasks." If a random sample of exercises is selected from a population of performance tasks, the student's percentage score can be considered an estimate of his or her proficiency (Kriewall, 1972). (An
assumption must hold: All exercises in the population are of equal difficulty for the student. This assumption appears to be more theoretical than practical.

Sensitivity to instruction seems to be moving to the forefront of criterion-referenced measurement's answer to norm-referenced item analysis procedures. Cox and Vargas (1966), in an investigation of several different discrimination indices, discussed an index of the ability for an item to discriminate between pre-training and post-training performance. Klein and Kosecoff (1973) described four methods of item analysis for criterion-referenced tests: (a) Comparison Group; (b) Single Group, Pre- and Post-test; (c) Single Group, Posttest Only; and (d) Single Group, Repeated Measures. The Comparison Group method discriminates between two groups, other things being equal. For example, one group is instructed and one group is not instructed with respect to an objective; or, two groups are used but requiring different levels of competency for the objective. The Single Group, Pre- and Posttest method is the case where a single group is tested before and after instruction on an objective. The Single Group, Posttest Only method involves a single group tested on an objective after a fixed period of instruction. If the time allotted is not sufficient for all to master the objective and if the students are heterogeneous with respect to ability, then classical item analysis methods, e.g., computing point biserials, may be used. The Single Group, Repeated Measures method involves giving the complete test repeatedly over time until the student masters the objective.
(Klein & Kosecoff, 1973, p. 7).

Each technique of item analysis described by Klein and Kosecoff (1973) has at least one of the following two basic concepts present in it:

1. An item is considered "good" if it is sensitive to instruction, that is, if performance on it is related to the degree of instruction obtained. (p. 6)

Usually the above construct is involved when there is little or no variation in test scores at a given testing. A problem is that such methods assume that the instruction was effective; and such methods also are likely to result in instructionally-dependent tests (p. 6).

2. An item is considered "good" if it discriminates between those who did well versus those who did poorly on the test as a whole or some "outside" criterion... (p. 7)

The above construct is related to classical item analysis approaches (p. 7).

A common way of expressing sensitivity of a test item to instruction is via a fourfold table:

<table>
<thead>
<tr>
<th></th>
<th>Fail</th>
<th>Pass</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass</td>
<td>A</td>
<td>B</td>
<td>A+B</td>
</tr>
<tr>
<td>Fail</td>
<td>C</td>
<td>D</td>
<td>C+D</td>
</tr>
<tr>
<td></td>
<td>A+C</td>
<td>B+D</td>
<td>N</td>
</tr>
</tbody>
</table>

The most important cell in the table is D, which indicates those who failed the item prior to instruction and who passed the item after instruction.
Various techniques are available to test the significance of the change from the first set to the second set of responses, such as chi-square tests (e.g., McNemar, 1962, pp. 52-56, 224-226). The fourfold table, though not always with chi-square tests advocated, seems to be a common way of representing the data for item sensitivity to instruction. For example, noted Raju (1974, p. 4), Cox and Vargas use (D - A); Roudabush uses D/ (D + C); Kosecoff and Klein use (D - C). But there are problems with interpretation of these tables. For example, Raju (1974) stated that a high value for $C$ "can be interpreted as indicating instructional deficiency; that is, ineffective instructional procedures" (p. 6). This writer does not completely agree; a high value for $C$ may also be indicative of a poor item. Likewise, a high value for $D$ may be indicative of either a good item to measure adequate instruction or the results of practice effect. The fourfold table is useful, no doubt, but this table cannot be considered criterion-referenced measurement's complete answer to norm-referenced measurement's item analysis. To carry the problem a step further, this fourfold table completely ignores the effect of the distractors of a multiple-choice item; are they good foils or aren't they?

It appears that much research is still necessary to find an acceptable item analysis procedure for criterion-referenced tests. Until then, while it is somewhat uncomfortable to this writer, we will have to heed the words of investigators such as Wilson (1974):

Test construction is not the strictly logical process that we might wish it to be. . . . Most of the really deep questions can only be answered by the exercise of well informed human judgment. (p. 36)
Reliability

Determining the accuracy of the test score (i.e., reliability or standard error of measurement) is one of the major problems in constructing achievement tests and analyzing and interpreting the results. Reliability is a concept which answers the question, "Does the test measure whatever it measures consistently?" According to Lord and Novick (1968):

The reliability of a test is defined as the squared correlation between observed score and true score. From the relation \( \rho_{xt} = \sigma_T^2 / \sigma_X^2 \), we see that the reliability of a test is a measure of the degree of true-score variation relative to observed-score variation. (p. 61)

These and other formulas in mental test theory are all based on one classic equation, the basic assumption in test theory:

\[ X = T + E \]

where \( X \) is the raw or observed score for a given person on a given test;

\( T \) is the true score for the same person on the same test;

\( E \) is the error component for the same person.

Three of the commonly-used techniques to assess reliability of norm-referenced tests are: (a) test-retest; (b) alternate forms; and (c) internal consistency. In the test-retest approach, the same test is administered twice to the same population, with, say, a two-week interval between testings. A Pearson product-moment correlation coefficient computed for the two testings defines the reliability coefficient. In alternate-forms reliability, two parallel forms of the same test are administered to the same population and a Pearson product-moment correlation coefficient is computed. Internal consistency measures answer a different question: "Do the items hang
together?" Internal consistency may be considered an indication of the precision of a test. Often it is computed by one of the Kuder-Richardson formulas which provide the average of all possible split-half reliabilities; the Kuder-Richardson formula gives a conservative (i.e., low) estimate of reliability.

The correlation coefficient needs variability of scores to provide a realistic estimate of linear association. As has been pointed out earlier in this chapter, a characteristic of a criterion-referenced test is that variability of test scores is irrelevant. Thus, it would appear that the traditional methods of assessing reliability are not appropriate for criterion-referenced tests.

As Boehm (1973) explained the problem:

The traditional questions of test reliability and validity are now being raised about criterion-referenced instruments. The inclusion of items on criterion-referenced measures which tap important components of a learning sequence (regardless of whether all pupils get them right or wrong) cuts down the variability among pupils, and can result in a restriction of the range of pupil scores. This restricted range can be inappropriately criticized from the point of view of norm-referenced measures, while appropriate measures of determining reliability and validity for criterion-referenced tests have yet to be developed. (p. 118)

Popham and Husek (1969) reported that a criterion-referenced test should be internally consistent to the extent that all items referenced to a criterion ought to be similar as to what they are measuring. They noted that if everyone obtained a perfect score on the test, a zero internal consistency estimate would result by traditional techniques. Popham and Husek pointed out that it is still possible to obtain a high average inter-item
correlation or a high test-retest correlation but that low correlations do not necessarily make for a poor test (p. 5).

Probably the statistic that has generated the most attention—and the most criticism—is Samuel Livingston's (1972a) reliability coefficient. Livingston, using assumptions from classical test theory, derived a coefficient based upon deviations of scores from the criterion score as opposed to deviations of scores from the mean. Where the reliability of a norm-referenced test is defined as the squared correlation between observed score and true score, Livingston defined reliability of a criterion-referenced test as the "squared criterion-referenced correlation between observed and true score" (p. 16). The major difference between the two measures is that in the derivation Livingston substituted mean squared deviations about the criterion score, $C_X$, instead of mean squared deviations about the mean, $\mu_X$.

The Livingston formula for reliability can be expressed as follows (Stanley, 1971, p. 435):

$$\rho_C^2(T_X, X) = \frac{\sigma_T^2 + (\mu_X - C_X)^2}{\sigma_X^2 + (\mu_X - C_X)^2}$$

where $\mu_X$ is the mean score and $C_X$ is the criterion score. From this formula, it can be seen that if the criterion score, $C_X$, is equal to the mean, $\mu_X$, then Livingston's coefficient is identical to the norm-referenced reliability coefficient, i.e., $\sigma_T^2 / \sigma_X^2$.

Harris (1972) presented an alternative interpretation to the Livingston (1972a) coefficient. Harris showed that Livingston's coefficient is
identical with a "conventional reliability coefficient when that coefficient is based on two [pooled] populations with means equally distant above and below C X" (p. 27). Harris also noted that one can artificially increase Livingston's coefficient by manipulating the criterion score. Finally, Harris demonstrated that because the standard error of measurement is not affected by the range of talent in a population, but Livingston's reliability coefficient is so affected, therefore Livingston's "larger coefficient does not imply a more dependable determination of whether or not a true score falls below (or exceeds) a given criterion value" (Harris, 1972, p. 29).

Hambleton and Novick (1973) suggested that:

Livingston misses the point for much of criterion-referenced testing. It is not, as he suggests, "to know how far (a student's) score deviates from a fixed standard." More typically we feel the problem is one of deciding whether a student's performance level is above or below some cutting score. (p. 168)

Raju (1973) stated that the larger reliability coefficients obtained by Livingston's (1972a) formula are "misleading indeed." Raju was able to derive the same formula as Livingston's by assuming that error variances are identical in two populations of interest and that the variance of the second population is greater than that of the first by an amount equal to \((\mu - C X)^2\). Raju concurred with Harris's (1972) conclusion that the larger reliability coefficients "do not imply a more dependable determination of whether a true score falls below a given criterion or not." Raju (1973) also showed that the assumption that the true score variance of the second population is greater than that of the first by \((\mu - C X)^2\) may be "impractical
and unreasonable."

Shavelson, Block, and Ravitch (1972, p. 136) suggested that Livingston's (1972a) reliability coefficient, because it is not a reliability coefficient in the traditional sense, should be given a different name.

In fairness to Livingston, it should be noted that he has responded to some of the criticism of his coefficient (Livingston, 1972b, 1972c, 1973), e.g., stating that "criterion-referenced test score interpretations do not require that the criterion score be conceptualized as the mean of some distribution" (Livingston, 1972b, p. 31).

Despite Livingston's attempts to defend his coefficient, there seems to be enough criticism of it to warrant looking toward something else. This writer prefers to look for an analogous concept of reliability rather than an analogous derivation of a reliability coefficient. Recall that reliability is a measure of consistency of scores. Perhaps, as Kriewall (1972) suggested, the analogy in criterion-referenced tests is that "reliability measures the extent to which a repeated measure would agree with the original measure on a group of examinees."

According to Harris (1974b), there are two questions one can ask about a mastery test:

1. How well does the test sort students into two groups?
2. How well does the test sort students into the correct two groups?

(p. 104)

Harris stated that these two questions correspond roughly to the questions of reliability and validity in more traditional contexts. Harris pointed out
that if a mastery test is valid, it will not only sort students into two categories but into the correct two categories. Harris derived an Index of Efficiency, an index of how well a mastery test sorts "defined samples of students into categories" (Harris, 1974b, p. 106). The derivation is done in the absence of criterion data; it assumes that the student produces (not chooses) a response, and assumes that there are $k$ items on the test. Drawing on Richardson (1936) --criterion of two categories, Fisher (1936) --linear discriminant function for two groups, and Tatsuoka (1971) --canonical correlation equivalents of discriminant functions, Harris (1974b, p. 107) derived the following formula:

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

where $SS_b$ and $SS_w$ refer to the sum of squares between and the sum of squares within for an analysis of variance of total scores on $k$ items for the two groups; $\mu_c^2$ is the squared canonical correlation which is also the squared Pearson product-moment correlation between total test score and a dummy variable designating the two-group sort. $\mu_c^2$ is also, thus, the squared point-biserial correlation coefficient. When $SS_w = 0$, $\mu_c^2$ can be +1.00 for two different total scores only. This situation would correspond to a phi coefficient of 1.00 (Harris, 1974b, p. 108).

Features of Harris's Index of Efficiency include:

1. [The coefficient] can be conceived as the ratio of true score variance to obtained score variance for a particular definition of true score.
2. The largest $\mu_c^2$ for a given test is an upper limit to the validity of
the mastery test when validity is measured in analogous form. (Harris, 1974b, p. 108)

Harris did not see, however, a way to determine a confidence interval for this index without using "possibly restrictive distributional assumptions" (p. 109).

A disadvantage of Harris's (1974b) index appears to be that the derivation assumes a produced response, as opposed to a multiple-choice response. Therefore, one could ask whether Harris's index is applicable to a multiple-choice criterion-referenced test. Also, Raju (1973) pointed out that Harris's index, like Livingston's coefficient, is influenced by the criterion or cutoff score, though the effect of that influence is not as well known at present. And, Harris himself advocated further research.

Another investigator, Jackson (1971), suggested that:

One way that "reliability" [of a criterion-referenced test] might be analyzed is through comparison of inferences made for a group of individuals on one form of a test with the inferences yielded by an alternate form developed independently with identical procedures. (p. 12)

Jackson recommended using an index of agreement between the two forms, perhaps a contingency coefficient. The disadvantage of Jackson's approach, of course, is that it would require alternate forms of a test. However, given the problems of other measures of reliability, one should not dismiss the basis of this idea too readily.

The ideas of Jackson (1971) and of Kriewall (1972) about reliability suggest perhaps that some sort of a fourfold table would have some potential in assessing reliability in the criterion-referenced approach. A method which might prove to have some promise is that advocated by Swaminathan,
Hambleton, and Algina (1974), who defined reliability as:

... a measure of agreement over and above that which can be expected by chance between the decisions made about examinee mastery states in repeated test administrations for each objective measured by the criterion-referenced test. (p. 263)

Using Cohen's (1960) kappa coefficient, this approach involves the entire decision-making process; that is, there must be a way to separate examinees into a mastery/nonmastery state (or several mastery states). It involves administering the same test twice and determining the consistency of placement. This method of reliability should be further explored, along with other measures that deal with consistency of sorting.

**Validity**

Determining the validity of test scores is another major problem in the construction of achievement tests and in the analysis and interpretation of the results. Validity is a concept which answers the question, "Does the test measure what it purports to measure?" A test is valid to the extent that it accurately measures the objectives for which it was designed. There are several methods of determining the validity of a test for a particular purpose:

**Content validity** is the extent to which test items appear to represent the objective given on the test. Often a number of experts are asked to judge the relevancy of the items to a set of objectives.

**Concurrent validity** is the extent to which scores on a test correlate with an external criterion when both measures are taken at approximately
the same point in time. Often test scores are correlated with other test scores or with teachers' judgment of the examinees' success.

**Predictive validity** is the extent to which scores accurately predict an external criterion. The same type of procedure is followed for predictive validity as in determining concurrent validity except that there is a time lapse between the time of testing and the measurement of the external criterion (e.g., several months to a year or more).

**Construct validity** is the extent to which a test measures a particular theory or construct on which the test itself was based. Often test scores are correlated with other test scores, some of which supposedly measuring the same construct, others not.

It should be pointed out that a test may be valid for one purpose and yet be invalid for another.

Popham and Husek (1969) observed that:

Many of the procedures for assessing the validity of norm-referenced tests are based on correlations and thus on variability. Hence, with validity, as with reliability, the results of the procedures are useful if they are positive, but not necessarily devastating if they are negative. (p. 6)

Much of the literature on the validity of a criterion-referenced test falls back on content validity—probably the easiest validity to measure because it tends to be based on expert judgment rather than on statistical technique. Lack of a statistical index does not make content validity a bad technique, however.

Jeter (1974), in developing a criterion-referenced test for Criminal
Investigation Division supervisor for the U.S. Army service schools, stated:

If the requirements are realistic, by reflecting the actual job requirements, and if the tests are taken directly from those objectives, then the tests must have face-, content- and predictive-validity. This was the logic used to justify the cost involved in administering the actual test instrument. (p. 32)

(This writer might not be so inclined to assume predictive validity from such a technique.)

Guzaitis (1973) also spoke to the issue of content validity for a criterion-referenced test:

Since criterion-referenced test items are evaluated for validity on their face, and the objectives they reflect are judged on their relation to the curriculum, there are fewer ways to evaluate these instruments than with the traditional norm-referenced variety. Therefore, careful attention to construction is even more crucial than it ever has been before.

Jackson (1971) referred to the "definitional validity" of a criterion-referenced test. What Jackson apparently was referring to is what others might call the content validity of a test; for a test to have definitional validity, the item generating rules must result in items that reflect the universe of content. Jackson felt that alternate forms would provide a check on validity, providing "cross-validation of the representativeness of the particular samples of tasks in each of the tests" (p. 12). This statement seems to be stretching the concept of cross-validation somewhat and may well be more indicative of alternate-form reliability.

Kriewall's (1972) item-sampling model "begins with the assumption of *prima facie* content validity," by assuming that a learning objective "is
defined by a specified item population." Kriewall stated, "In the final
analysis, the test builder must make subjective decisions concerning a
given item's relation to whatever it is he wants to measure."

Womer's (1970) description of content validity and item analysis for
NAEP further labors this point:

With this type of reporting in mind [i.e., reporting only p-values
for each item alternative], National Assessment developed its exercises
with an eye to content validity, as judged by subject matter specialists,
other educators, and laymen. The exercises were not item analyzed
(there is no total score) nor were they related to future performance
(there are no criterion measures). The purpose of National Assess­
ment exercises in toto is to describe, by example, what most young
people know and can do, what about half can do and what very few can
do. The purpose of a single exercise is to stand as one example of a
meaningful knowledge or skill or attitude that relates to a specific ob­
jective in a given subject area. (p. 2)

In a later paper on National Assessment, Wilson (1974) listed two
major concerns of National Assessment as the content validity and impor­
tance of assessment items:

1. Is this exercise a valid measure of the objective for which it was
   written?
2. If it is valid, is it an important or a trivial measure of the objec­
tive? (p. 36)

For the first question, on content validity, Wilson (1974) reported that
National Assessment uses human judgment; for the second question, on
importance, judgment of subject matter experts is used (p. 36). Wilson
also noted that for some of the National Assessment test items:

... another measure of their validity can be obtained by examining
the assessment response data. If an item is administered to two groups,
one of which has had no training, the results can be viewed as one
measure of the item's validity. In the ideal case, a valid item would...
yield a score near zero for the untrained group and approach 100% correct for the highly trained group. (p. 36)

Note that Wilson's (1974) approach to validity is merely a sensitivity to instruction approach advocated by some as a method of criterion-referenced item analysis. In view of the nature of criterion-referenced measurement versus norm-referenced measurement, however, sensitivity to instruction, along with content validity, may well serve the function for both item analysis and validity of a criterion-referenced measure. In fact, this concept may even be extended to include reliability. The same paradigm may be used to assess reliability as item analysis and validity if a test-retest or alternate-form procedure is utilized.

There are other approaches to criterion-referenced validity that ought to be examined. The fourfold table, which seems to be popular among criterion-referenced advocates, is one approach of Harris's (1974b) to assess the validity of a mastery test:

For a mastery test the ultimate validity question is the question of the extent to which the test sorts students into the correct two categories. Given an appropriate criterion, it is possible to develop the two-by-two table that results from classifying students as "true masters" or "true non-masters" on the basis of the criterion data and simultaneously classifying them as "indicated masters" or "indicated non-masters" on the basis of the mastery test. An appropriate interpretation of these data provides a validity statement for the test. (p. 109)

Harris noted that one can use either a phi coefficient or a tetrachoric coefficient to summarize a two-by-two table but that "not all 'experts' agree on which, if either of these two, should be used" (p. 112).

Kifer and Bramble (1974) also presented a fourfold table, or loss matrix, with the test decision on one axis, the true state on the other.
It should be noted that there are investigators who do not treat criterion-referenced measures differently from norm-referenced measures, at least in terms of some of the concepts. Young, Regedal, and Knapp (1973), for example, noted that:

Although the shift in emphasis from norm-referenced to criterion-referenced measurement has resulted in refinements of the statistical techniques used to analyze the tests . . . , in most situations standard statistical techniques may be applied in analyzing results. (p. 909)

The above statement may be open to some question; but nevertheless the authors correlated their criterion-referenced measure (Tests of Achievement in Basic Skills [TABS], Math, Level B) with teachers' marks. In a predictive validity study, TABS was administered in the fall of a school year to grades 4-6; course marks were assigned in the spring, without the teachers having access to the test scores. Correlations of the total TABS score with teachers' marks were .43 (grade 4; N = 198), .62 (grade 5; N = 316), and .73 (grade 6; N = 245). The authors considered these correlations "consistent with results obtained in previous studies relating test scores to achievement as reflected in teachers' marks" (p. 911). While these results are interesting, they are not necessarily indicative of the success of using norm-referenced indices for criterion-referenced tests. In this study, total test score, apparently measuring a number of mathematics objectives, was used; the consequence might well result in the equivalent of a typical achievement test with typical variation of test scores. Also, one could question the relevancy of performing a predictive validity study of a criterion-referenced test; it would seem that a criterion-
referenced test, whether diagnostic or mastery, would have a purpose of determining the current level of performance of the students taking the test.

But Klein and Kosecoff (1973) provided additional support that perhaps some investigators are abandoning the criterion-referenced measure's lack-of-variance philosophy:

The . . . problem [of no variance on the test scores] . . . usually appears to be more theoretical than actual, because students do vary in their performance. This variation may be due to a number of factors including the students' general intellectual ability, cultural and environmental backgrounds, and the quality of instruction they receive. . . . Reports of "no variance" usually stems from failure to sample enough students and/or from the failure to examine the rate at which students master items and objectives. . . . The real problem, therefore, is not in finding variance but in identifying just that portion of the variance that is due to the student's degree of mastery of the particular objective on which the CRT is based rather than variance due to some extraneous influence. (p. 8)

This writer is beginning to see a number of parallels with criterion-referenced testing and with industrial or personnel testing; these parallels will be more fully discussed in Chapter V. For example, the concept of synthetic validity seems related to that of content validity of criterion-referenced measures. In a review of methodology and technology of educational and psychological testing, Mayo (1968) noted:

Lawshe's development of synthetic validity (Lawshe and Balma, 1966) tends to overcome the limitations of predictive and concurrent test validation studies as they are frequently performed in industry. The concept was well defined by Balma as the "inferring of validity in a specific situation from a logical analysis of jobs into their elements, a determination of test validity for these elements, and a combination of elemental validities into a whole." (Mayo, 1968, p. 96)

Content validity, sensitivity to instruction, and relationship of the
decision of the test to that of expert judgment seem to be at present the
most viable approaches to assessing the validity of the criterion-referenced
test.

**Mastery Criterion**

One of the most important problems, perhaps the most important
problem, is, ironically, defining mastery and determining how many test
items are required to establish mastery of an objective. Most systems in
use are purely arbitrary; for example, 80% seems to be a popular figure.
In general, percent of items correct to establish mastery seems to range
around 70% to 90%. Perhaps these percentages stem from Bloom's (1968)
article on "Learning for Mastery" where he stated that:

Most students (perhaps over 90 percent) can master what we have
to teach them, and it is the task of instruction to find the means
which will enable our students to master the subject under considera-
tion. (p. 1)

Unfortunately, few readers seem to notice that in Bloom's next sentence
he added:

Our basic task is to determine what we mean by mastery of the subject.
\ldots (p. 1)

Thus, it would appear that Bloom has said that perhaps 90% of students can
master something, but we must define what mastery means. The definition
seems somewhat circular.

Skager (1974, p. 56) stated that the issue of defining mastery is
related to reliability. There must be enough items such that the test is
reliable enough to state that there is mastery. Skager added that fewer
items would be needed if guessing could be ruled out. Of course, guessing cannot be ruled out for any multiple-choice test.

Klein and Kosecoff (1973) restated the problem and provided some additional elaboration on it:

There is rarely a good way of defining exactly what is meant by "mastery." Arbitrary definitions, such as 85 percent correct, are rampant; but there is rarely any satisfactory criterion for setting such standards of performance. Further, a mastery score often hides the true level of student performance. In other words, if the student failed to achieve mastery did he miss by a little or miss by a great deal; or if he made it, did he just squeak by? (p. 9)

Kifer and Bramble (1974) pointed out yet another problem with defining mastery:

Logically, a criterion-referenced test should be "graded" dichotomously--a person either gets all of the items correct or he does not. By establishing an 80% criterion level, for example, the test maker is saying implicitly that it does not matter which 80% of the items the person gets correct. But yet, some items may measure more important performance than others. . . .

Davis and Diamond (1974) provided some additional insight:

Strictly speaking, mastery is defined as complete knowledge, skill, or control; so "partial mastery" is as self-contradictory a phrase as "partial uniqueness." The term "mastery," therefore, should be used to describe the status of only those examinees who, it may be inferred, can mark correctly all items in the population of which the subset that makes up a criterion-referenced test is a representative sample. (p. 133)

Davis and Diamond suggested setting the cutting score at a point lower than mastery but high enough to meet practical considerations. They presented tables to show probabilities of specified competence levels given only the examinee's test score--via Bayes' theorem. The authors pointed out that some subject areas require greater competence levels. For example,
mathematics would require greater competence levels than, say, social studies; mathematics skills tend to build on previous mathematics skills (p. 134).

According to Harris (1974b), the proper function of a mastery test is to provide the means for "making a 'mastery'-'non-mastery' decision for a given student" (p. 99). Harris operationally defined the term, concept of mastery, as "a proportion of a population of items that a properly instructed student should be able to answer correctly" (p. 105). Harris stated that it is possible to set two cutting scores—one, above which the student is considered to have mastered the instruction, a second, below which, the student is considered to be a nonmaster. The student who scores in between these two cutting scores is considered to be "in limbo" (p. 105).

Harris (1974b, p. 105) showed that Wald's (1947) sequential procedure can be adapted for a fixed length mastery test. Given (a) number of items, (b) percent of items for mastery, (c) percent of items for non-mastery, (d) probability of incorrect decision of mastery, and (e) probability of incorrect decision of nonmastery, the formula yields (a) mastery score or index (percent of items) and (b) nonmastery index (percent of items). This approach seems somewhat circular to this writer, however, for one must input what he wants to be the percent of items correct for mastery and nonmastery. Also, probabilities of incorrect decisions must be estimated to input into the formula. Another problem with this approach,
according to Harris, is that application of Wald's sequential procedure demonstrates that "very short tests often do not lead to informed decisions."

Harris, however, prefers application of Wald's approach to the approach presented by Millman (1973) "since Millman's work solves the 'wrong' problem" (Harris, 1974b, p. 105). Millman used a binomial-based model to determine passing scores for tests of varying lengths.

Luebke (1972) presented the state of the art in 1972 which is not so untrue today:

Evaluation by means of criterion-referenced measurement necessitates the establishment of an acceptable level of performance or criterion. In all the literature on criterion-referenced measurement, the authors have stated that the student must achieve the criterion score. . . . To date, however, no one has outlined a systematic approach for determining a realistic criterion. . . . (p. 1)

To determine the criterion score on a dental pharmacology test, Luebke had three subject matter experts estimate the percentage of the students that would answer each test item correctly (p. 2). The sum of these "percentage scores" would correspond to the mean test score (expected) of the student group. After applying the Spearman-Brown formula, the inter-rater reliability for the three judges was .63. (This reliability figure does not seem particularly high to this writer.) Comparison of criterion scores with mean student performance showed a close correspondence (p. 3). (The results either mean that the judges can set satisfactory criterion scores or that the judges can predict fairly well the existing level of performance on such examinations.)

Meskauskas (1976) wrote a very comprehensive review of evaluation
models for criterion-referenced testing. According to Meskauskas, mastery models can be placed into two broad models: (a) continuum models—"mastery as an area on a continuum"; and (b) state models—"mastery as all-or-none" (p. 134).

Characteristics of continuum models of mastery include:

1. Mastery is viewed as a continuously-distributed ability or set of abilities.
2. An area is identified at the upper end of the continuum, and if an individual equals or exceeds the lower bound of this area, he is termed a master.
3. The goal of measurement is to obtain information for the purposes of educational decision-making, which explicitly follows the classification decision. (p. 134)

Meskauskas (1976) noted variations on the above: "Some writers have viewed mastery in terms of a continuum of skill ranging from none to perfection" (p. 135). Examples of continuum models of mastery are: (a) Nedelsky's Minimum Pass Level (MPL) Method; (b) Ebel's Method of Passing Score Estimation; and (c) Kriewall's Binomial-based Model (pp. 135-141).

Characteristics of state models of mastery include:

1. Criterion-referenced test (CRT) true-score performance is viewed as an all-or-none dichotomous task.
2. The standards or cutting score that should be used in an error-free situation is implied as part of the model.
3. Considerations of measurement error essentially always result in the adoption of standards that demand less than the model seeks. (p. 142)

Examples of state models of mastery are: (a) Emrick's Mastery Testing Evaluation Model; and (b) Roudabush's Dichotomous True-Score Models (pp. 143-148).
Meskauskas also discussed several "mixed mastery models": (a) Millman's Binomial-based Decision Model; (b) the Davis and Diamond Bayesian Model; and (c) the work of Novick and collaborators (pp. 148-155).

In his discussion, Meskauskas noted that the models presented "need to be validated to provide users with data on which to make choices" (p. 216). He also noted:

The State model may appear, at first glance, to represent an unreasonable approach to learning and evaluation. Perfection often appears to be something to strive for, but not to reach. And yet a great deal of what is learned, particularly in situations where errorless replication will be required, follows this model. (p. 216)

It is difficult to accept the philosophical ramifications of Meskauskas's (1976) statement—at least based on current criterion-referenced test technology. Though in theory mastery should mean perfection, the state of the art does not provide for a measure of perfection. Harris (1974b) stated that "one argument or position rests on the truism that no individual ever 'masters' a subject matter or an art, and concludes that a mastery test is a contradiction in terms" (p. 99). While the above position is not Harris's own view, it does come close to the view of this writer who suspects that "mastery" may be a misnomer. Mastery, as perfection, is a concept which nobody really can attain; "mastery" should be viewed as a continuum representing degree of knowledge or competence. The criterion level, which psychometricians are trying so hard to define, should be considered minimum acceptable performance, not mastery.
An earlier section of this chapter referred to the work of Raju and associates (SRA, 1975) in choosing three items as the ideal number of test items to measure an objective. In doing so, Raju also recommended a criterion for mastery of three out of three, based on the probability of attaining mastery by chance alone. Raju reasoned as follows:

Any criterion for mastery of learning objectives involves two factors: the number of items that measure an objective and the number of items that must be answered correctly to attain mastery. [Science Research Associates'] Mastery tests contain three items to measure each objective. The student must correctly answer all three items to show mastery. The criterion was decided on after careful consideration of the chance factor in attaining mastery. The lengths of the tests were also taken into account. . . . (SRA, 1975, p. 5)

Table 2 is reproduced from the referenced publication. It shows the "probabilities of attaining mastery by chance alone when each objective is measured by one, two, three, four, or five items and when the criterion for mastery varies."

The probability of attaining mastery by chance alone when the criterion is 3 out of 3 is .016 [with each item having four choices]. The probabilities are lower only for the criteria of 4 out of 4 and 5 out of 5. The probability is the same for 4 out of 5 as it is for 3 out of 3. The probability of .016 is low enough so that a statement of mastery using 3 out of 3 has practical significance. Since a test with four or five items measuring each objective could become unwieldy in length, the criterion for mastery was set at 3 out of 3. (SRA, 1975, p. 5)

The Raju approach is not the final answer, of course. As any teacher or test maker knows, a criterion of three out of three is as easy or as difficult as the items and their distractors. Nevertheless, Raju's approach is simple and avoids the many assumptions often needed to apply the numerous mathematical models appearing in the literature.
TABLE 2

PROBABILITY OF ATTAINING MASTERY BY CHANCE ALONE

<table>
<thead>
<tr>
<th>Mastery Criterion</th>
<th>Total Number of Items per Objective (Each item has four choices)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1 or more correct</td>
<td>0.250</td>
</tr>
<tr>
<td>2 or more correct</td>
<td>---</td>
</tr>
<tr>
<td>3 or more correct</td>
<td>---</td>
</tr>
<tr>
<td>4 or more correct</td>
<td>---</td>
</tr>
<tr>
<td>5 correct</td>
<td>---</td>
</tr>
</tbody>
</table>

Norm-Referenced Criterion-Referenced Measures

Guzaitis (1973) traced some attempts by educators to obtain criterion-referenced-type information from norm-referenced tests via item response analysis, i.e., examining a student's response to standardized test items:

On the face of it, this seemed like an efficient way to proceed. By gathering diagnostic and mastery information for the teacher, while providing summary information for the administrator, it would have seemed that we were able to get our two birds with one stone. For a while it didn't seem to matter that the slingshot was unwieldy and the birds hard to retrieve.

Well, as we know, criterion-referenced tests by themselves became very popular both as replacements to and more usually as supplements to norm-referenced standardized achievement tests. It almost seems ironic that where before educators were seeking criterion-referenced data from norm-referenced tests, a movement now is to obtain norm-referenced data from criterion-referenced tests. The movement makes some sense but does have some definite pitfalls.

Womer (1970) stressed that there are no norms or standards against which to compare the results of NAEP testing. If two-thirds of the 17-year-olds can answer a particular science item, is that good or bad, asked Womer. Womer suggested that it must be a "personal, thoughtful judgment" of whether the students are learning what they "should" learn (p. 2).

Klein and Kosecoff (1973) noted that though the primary orientation of criterion-referenced testing is to indicate what the student can do, the information of which objective a student has mastered might be met with, "Is this performance satisfactory?" In other words, though the emphasis
on criterion-referenced tests is on describing behavior in the absolute and not the relative, a normative framework seems desirable and can be provided. Such information that can be provided includes:

1. The number or percent correct on a given objective or set of items than [sic] encompass a few highly related objectives.
2. "Mastery" of a given objective or set of items where "mastery" is defined in terms of a certain level of performance such as 90 percent correct.
3. The time it takes (such as class hours or calendar days) for an individual to achieve a given performance level (including what has been defined as "mastery").
4. The time (in minutes or hours) it takes a student to perform a certain task or set of tasks related to an objective (such as correctly computing the product of all single digit numerals).
5. The probability that the student is ready to begin the next level of instruction (this may be based on both the number of items correct and the pattern of answers given to these items).
6. The percentage of students who "pass" each item; that is, the item's difficulty. This kind of score is used exclusively in program evaluation where each item or task is considered important in itself. (Klein & Kosecoff, 1973, p. 8)

Kifer and Bramble (1974) stated:

We believe that in practice it is impractical, and perhaps impossible, to separate criterion-referenced measurements from a normative rubric. In a fundamental sense the test maker imposes a normative framework on his criterion-referenced test when he defines a domain of items or decides which testing objectives should be excluded or included in the development stage of the test... and when persons set a criterion level.

Finally, Farr and Roser (1974, p. 597) advocated the use of what they call "standardized criterion tests," tests which would provide both criterion scores and normative data. And Popham (1976) recommended normative data for criterion-referenced tests.

Popham (1976) observed that:
... educators who use criterion-referenced tests for evaluation purposes are beginning to encounter a troublesome problem. They can describe with clarity what students can do, for example, at the close of an instructional program. Nevertheless, people want to know how well the students should be doing. (p. 593)

Popham argued that fears that providing norm-referenced data for criterion-referenced tests will turn criterion-referenced tests back into norm-referenced tests are unfounded. What Popham and others are advocating is maintaining the information that can be provided by criterion-referenced tests but providing in addition information which will tell how well the group is doing.

In theory, these educators have an excellent point. It is really not enough to say that 64% of the class has mastered a particular learning objective. The follow-up always is, "How good is this performance?" Unfortunately, to provide a reliable and valid measure of this goodness requires the same standardization process that a norm-referenced achievement test has. Standardizations if done properly are expensive and time consuming. And there are many, many more learning objectives that would have to be standardized for a criterion-referenced test; a norm-referenced test is only a sampling of various learning objectives for a particular subject matter; criterion-referenced tests are more thorough in their coverage. Who is to decide which objectives will be covered and standardized? How can a national, or regional, standardization be logistically accomplished? There are many problems with this approach and perhaps only compromises can be accomplished—such as test publishers
providing both norm-referenced and criterion-referenced data on a limited number of objectives that comprise a standardized achievement test battery. This issue will be an area well worth watching in the very near future.

Construction of a Diagnostic-Prescriptive System

For Developing Measurement Competency for Prospective Teachers

This section describes application of criterion-referenced procedures to the construction of a diagnostic-prescriptive system for developing measurement competency for prospective teachers. The work described in this section has as its origins Mayo's (1967) report on Pre-Service Preparation of Teachers in Educational Measurement. Mayo, after reviewing the few studies done in the area of pre-service preparation of teachers in measurement, had two conclusions:

1. There was a dearth of systematic and effective preparation of teachers in measurement; and
2. In-service teachers felt strongly their need for competency in measurement and evaluation. (p. 3)

With a view toward surveying and upgrading measurement competencies of beginning teachers, Mayo identified the set of competencies that would be needed by beginning teachers and constructed an objective test to measure these competencies.

From a subject-matter outline developed by a National Council on Measurement in Education (NCME) Committee on Pre-Service Preparation of Teachers in Measurement Competency, Mayo's project produced, after several stages, a 70-statement Checklist of Measurement Competencies.
(See Appendix A.) The Checklist, like the subject-matter outline which preceded it, consists of four content categories of competencies: (a) standardized tests; (b) construction and evaluation of classroom tests; (c) uses of measurement and evaluation; and (d) statistical concepts. Relative importance of the various competencies was obtained by Mayo from a national sample of teachers, principals and superintendents, college professors, measurement specialists, and miscellaneous personnel (primarily counselors and school psychologists). Using the ratings of relative importance of the Checklist statements as a guide, Mayo developed two forms of a Measurement Competency Test, a 60-item objective test to measure the content and behavior of the Checklist statements. Mayo's (1967) Measurement Competency Test appears in Appendix B.

A logical extension of Mayo's work would be the construction of a diagnostic-prescriptive system keyed to Mayo's Checklist statements.

The first step in the construction of a diagnostic-prescriptive system for developing measurement competency for prospective teachers was the identification and definition of the specific skills to be mastered in the area of measurement competency for prospective teachers. In order to correspond with the work previously done in this area by Mayo (1967), the four content categories—i.e., (a) standardized tests, (b) construction and evaluation of classroom tests, (c) uses of measurement and evaluation, and (d) statistical concepts—of Mayo's (1967) Checklist of Measurement Competencies and their corresponding Checklist statements were used as
the basis of the skills to be evaluated. Investigation of the content of several current textbooks in tests and measurements verified the appropriateness of using this Checklist.

For each Checklist statement, one or more test items were written to measure the indicated content and behavior. The number of items written per objective depended on this writer's judgment as to how many items were sufficient to properly cover each objective. For example, only one item ("What are some advantages and disadvantages of standardized tests?") was deemed necessary to measure Checklist statement number 1 ("Knowledge of advantages and disadvantages of standardized tests"), while four items were constructed to measure Checklist statement number 5 ("Knowledge of sources of information about standardized tests").

Several sources contributed to the writing of the diagnostic test items: (a) instructor's manuals for various statistics and tests and measurements textbooks, especially Thorndike's and Hagen's (1969) *Measurement and Evaluation in Psychology and Education* (3rd edition), Gronlund's (1971) *Measurement and Evaluation in Teaching* (2nd edition), Noll's and Scannell's (1972) *Introduction to Educational Measurement* (3rd edition), Stodola's and Stordahl's (1967) *Basic Educational Tests and Measurements*, and Minium's (1970) *Statistical Reasoning in Psychology and Education*; (b) examination items from courses in statistics and tests and measurements taught by Mayo and by this writer; and (c) original test items written by this writer. One advantage in using these sources
was that generally the items had already been tried out in the classroom situation and presumably only the "good" items had been retained by the authors.

Following the construction of the four content categories diagnostic tests, a mapping of the learning objectives to specific pages in a major tests and measurements textbook was accomplished. Because the field try-out involved a class using Stanley's and Hopkins's (1972) *Educational and Psychological Measurement and Evaluation*, the objectives were mapped to pages in that textbook and were provided on *Prescription Sheets* for each student.

The complete list of components of the Diagnostic-Prescriptive System for Developing Measurement Competency for Prospective Teachers is as follows:

1. **Diagnostic Tests.** Four diagnostic tests, entitled *Diagnosis of Measurement Competency* and corresponding to Mayo's (1967) four content categories of measurement competencies, were constructed. Part I (Standardized Tests) covers 10 learning objectives with 31 test items; Part II (Construction and Evaluation of Classroom Tests) covers 13 learning objectives with 44 test items; Part III (Uses of Measurement and Evaluation) covers 13 learning objectives with 38 test items; Part IV (Statistical Concepts) covers 34 learning objectives with 63 test items. (See Appendix C.) All four diagnostic tests cover a total of 70 learning objectives with 176 test items.
2. **Answer Keys.** Each Diagnosis of Measurement Competency test has a corresponding **Answer Key.** Users of the system are instructed to circle the item numbers of every incorrect answer and to refer to the **Prescription Sheet** to identify the corresponding learning objectives and the prescriptive page references. (See Appendix D.)

3. **Prescription Sheets.** The **Prescription Sheets** contain the prescriptive page references in Stanley's and Hopkins's (1972) textbook for each of the learning objectives covered by the diagnostic tests. Users of the system are instructed to consult the prescriptive page references in order to find material for remediation of the nonmastered learning objectives. (See Appendix E.)

4. **Instructor's Guide.** A brief **Instructor's Guide** explains the system to the user and provides suggestions for a **Progress Chart** to assist the tests and measurements instructor in following the progress of each student through the program. (See Appendix F.)

5. **Entry Survey Test.** In order to determine which of the four diagnostic tests would be needed by an individual student at some later time, Mayo's (1967) **Measurement Competency Test** can be administered as a general survey instrument (optional) and item response data analyzed. Table 3 presents the item numbers of the **Measurement Competency Test** which correspond to the four diagnostic tests used in this study. If, for example, the student misses several MCT Form A items from questions 46-60, the student would be directed to Part IV of
TABLE 3

ITEM NUMBERS OF MEASUREMENT COMPETENCY TEST
CORRESPONDING TO CONTENT CATEGORIES

<table>
<thead>
<tr>
<th>Content Categories</th>
<th>Item Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Form A</td>
</tr>
<tr>
<td>I. Standardized Tests</td>
<td>1-15</td>
</tr>
<tr>
<td>II. Construction and Evaluation</td>
<td></td>
</tr>
<tr>
<td>of Classroom Tests</td>
<td>16-30</td>
</tr>
<tr>
<td>III. Uses of Measurement and Evaluation</td>
<td>31-45</td>
</tr>
<tr>
<td>IV. Statistical Concepts</td>
<td>46-60</td>
</tr>
</tbody>
</table>

Diagnosis of Measurement Competency in order to ascertain which specific Statistical Concepts learning objectives the student is weak in.
CHAPTER IV

TRYOUT OF DIAGNOSTIC-PRESCRIPTIVE SYSTEM

The Diagnostic-Prescriptive System for Developing Measurement Competency for Prospective Teachers was tried out in a classroom environment in order to determine that the system could be easily incorporated into the class operations; that is, the classroom management aspects of the system were investigated.

The system was tried out in an undergraduate tests and measurements course at Loyola University of Chicago. The textbook in use was Stanley's and Hopkins's (1972) Educational and Psychological Measurement and Evaluation. The objectives of the course were organized by the instructor in such a way that the order of the four diagnostic tests was as follows:

1. Part III: Uses of Measurement and Evaluation
2. Part IV: Statistical Concepts
3. Part II: Construction and Evaluation of Classroom Tests
4. Part I: Standardized Tests

Subjects

Subjects were 22 students enrolled in an undergraduate tests and measurements course. The subjects participated in the study on a voluntary basis on their own time; and degree of involvement in the study did
not affect their course grade. However, it was announced by the instructor that some test items for the final examination of the course would be selected from items appearing in the system.

**Procedure**

During the course of the semester, the Diagnostic-Prescriptive System was made available to the students at the appropriate times during the course, generally upon completion of a particular unit of instruction. For example, Part IV: Statistical Concepts was made available to the students following the instructional segment on that topic. The students were instructed to take the diagnostic test on their own time, score their tests with the provided Answer Key, identify which learning objective they had not yet mastered by using the Prescription Sheet, and consult appropriate pages in their textbook by using the Prescriptive Sheet. Informal feedback about the use of the system was encouraged but not required.

**Ancillary Study**

Though the primary purpose of this classroom study was to try out the diagnostic-prescriptive system rather than to conduct an experiment, ancillary data were, nevertheless, collected.

In order to obtain some measure of pre-to-post gains, Mayo's (1967) Measurement Competency Test, Form A, was administered to each student in the class during the first week of instruction. The test was readministered during the final week of instruction, some 3.5 months after the first
administration. Scores on the test were not made available to the instructor and did not affect course grade.

In order to provide a baseline for the pre-to-post gains, data from the previous semester of a tests and measurements class (N = 19) taught by the same instructor with the identical syllabus were used. In this class, the diagnostic-prescriptive system was not present but pre and post Measurement Competency Test, Form A, scores were available.

By using the two tests and measurements classes in this manner, a modified Nonequivalent Control Group Design (Campbell & Stanley, 1963) could be employed. This design, common in educational research, involves giving both an experimental group and a control group a pretest and a posttest; but, in this design, the two groups already exist and are not formed by the random assignment of subjects from a population. In the true Nonequivalent Control Group Design, the assignment of the treatment (i.e., the diagnostic-prescriptive system) is "assumed to be random and under the experimenter's control" (Campbell & Stanley, 1963, p. 47). However, in this study, assignment of the treatment was not random; assignment was determined by the investigator.

Two hypotheses were postulated for this ancillary study: (a) There will be a significant improvement in scores on the Measurement Competency Test after using the diagnostic-prescriptive system; and (b) there will be a greater increase in scores on the Measurement Competency Test for a class using the diagnostic-prescriptive system over one which does not.
Results

This section describes the results of the application of criterion-referenced procedures to the construction and field tryout of a diagnostic-prescriptive system for developing measurement competency for prospective teachers.

The diagnostic-prescriptive system was used by the tests and measurements class for the entire semester. According to the instructor, the students did participate and neither he nor they reported any problems with the system.

The pre- and post-administrations of the Measurement Competency Test (Form A) were scored by hand twice in order to ensure accuracy of scoring. There were 18 students with complete data for the two administrations in the experimental group. The means and standard deviations of the two administrations are presented in Table 4.

The data indicated that there was a mean gain from pretest to posttest of 5.72 points. In order to determine whether this gain was significant, a one-tailed t-test was performed on the pre and post difference scores. The analysis indicated that there was a significant pre-to-post gain in scores, \( t(17) = 3.52, p < .005 \).

This analysis, however, does not separate the effects of instruction from the effects of the diagnostic-prescriptive system. In order to assist in this determination, data from the previous semester of a tests and measurements class taught by the same instructor with the identical syllabus
TABLE 4

MEANS AND STANDARD DEVIATIONS FOR MEASUREMENT COMPETENCY TEST (FORM A)

EXPERIMENTAL GROUP

<table>
<thead>
<tr>
<th>Measurement Competency Test (Form A)</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>23.17</td>
<td>28.89</td>
<td>+5.72</td>
</tr>
<tr>
<td>SD</td>
<td>6.41</td>
<td>8.61</td>
<td>6.90</td>
</tr>
</tbody>
</table>

Note. N = 18.
were used. In this "control" class, data on the Measurement Competency Test (Form A) had also been obtained pre and post by the instructor, without the intervening treatment of the diagnostic-prescriptive system. There were 15 students with complete data for the two administrations. The means and standard deviations for this control class are presented in Table 5.

The data of the control group were also analyzed for significance of the gain from pretest to posttest. A one-tailed $t$ test of the pre and post difference scores indicated that the mean gain of 4.53 was significantly different from zero, $t(14) = 4.07$, $p < .005$.

A third analysis compared the adjusted posttest scores of the treatment group versus those of the control group. The posttest scores were adjusted by means of an analysis of covariance technique (Winer, 1971). Analysis of covariance is a statistical technique used to control experimental error. In the present study, it was not possible to randomly assign students to the two groups (i.e., treatment and control); analysis of covariance is a technique that can be used to statistically adjust the posttest variable (the "criterion") by examining the regression of the criterion on the covariate (in this case, the pretest variable). It was hypothesized that there would be a greater amount of pre-to-post gain in the treatment group than in the control group. The results of this analysis of covariance are presented in Table 6. The analysis indicated that although the trend was in the hypothesized direction, the trend failed to achieve an acceptable
### TABLE 5

**MEANS AND STANDARD DEVIATIONS FOR MEASUREMENT COMPETENCY TEST (FORM A)**

**CONTROL GROUP**

<table>
<thead>
<tr>
<th>Measurement Competency Test (Form A)</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>25.00</td>
<td>29.53</td>
<td>+4.53</td>
</tr>
<tr>
<td>SD</td>
<td>4.56</td>
<td>4.86</td>
<td>4.32</td>
</tr>
</tbody>
</table>

*Note. N = 15.*
## TABLE 6

ANALYSIS OF COVARIANCE OF MCT POSTTEST SCORES

### DATA SUMMARY

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Adjusted Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>15</td>
<td>25.00</td>
<td>4.56</td>
<td>29.53</td>
<td>4.86</td>
<td>28.77</td>
</tr>
<tr>
<td>Treatment</td>
<td>18</td>
<td>23.17</td>
<td>6.41</td>
<td>28.89</td>
<td>8.61</td>
<td>29.52</td>
</tr>
</tbody>
</table>

F Statistic for homogeneity of within-class regression = .247.

### ANALYSIS OF COVARIANCE

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>4.51</td>
<td>1</td>
<td>4.51</td>
<td>.126*</td>
</tr>
<tr>
<td>Error</td>
<td>1077.77</td>
<td>30</td>
<td>35.93</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1082.28</td>
<td>31</td>
<td>34.91</td>
<td></td>
</tr>
</tbody>
</table>

*p = .726*
level of statistical significance, $F(1, 30) = .126$, $p = .726$.

Examination of the data revealed that four students in the treatment group exhibited zero or negative gain from pretest to posttest. This finding seemed difficult to explain; it would seem that a semester's instruction and/or the diagnostic-prescriptive system and/or maturation alone ought to account for some gain between the two testings. Discussion with the class instructor revealed that the student with the most negative gain ($-11$) was considered by the instructor as a "problem" student. Thus, the analysis of covariance was repeated with the data for this student eliminated. The results of this analysis are presented in Table 7. While the trend reached a lower level of significance, the trend still failed to achieve a generally acceptable level of significance, $F(1, 29) = .868$, $p = .359$.

Finally, as a matter of interest, the analysis of covariance was repeated after dropping the other three students who had experienced zero or negative gain (0, -1, and -2, respectively). The results, presented in Table 8, were borderline significant, $F(1, 26) = 3.456$, $p = .074$.

Another comparison which could be made is to Mayo's (1967) reported pre-to-post follow-up gains. Of 1,780 subjects who had taken the Measurement Competency Test (Form A), Mayo obtained retest data two years later on a follow-up subsample of 341 subjects, mean gain = 2.06, SD = 6.23. Using this group as a sort of conservative "control" group, the mean gains observed in the present dissertation study were compared. Both the dissertation treatment group and control group showed significantly higher
TABLE 7

ANALYSIS OF COVARIANCE
OF MCT POSTTEST SCORES
ELIMINATING ONE OUTLIER

<table>
<thead>
<tr>
<th>Group</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Control</td>
<td>25.00</td>
<td>4.56</td>
<td>29.53</td>
</tr>
<tr>
<td>Treatment</td>
<td>23.18</td>
<td>6.59</td>
<td>29.88</td>
</tr>
</tbody>
</table>

F Statistic for homogeneity of within-class regression = .318.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>23.35</td>
<td>1</td>
<td>23.35</td>
<td>.868*</td>
</tr>
<tr>
<td>Error</td>
<td>780.27</td>
<td>29</td>
<td>26.91</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>803.62</td>
<td>30</td>
<td>26.79</td>
<td></td>
</tr>
</tbody>
</table>

*p = .359
TABLE 8

ANALYSIS OF COVARIANCE
OF MCT POSTTEST SCORES
ELIMINATING FOUR OUTLIERS

---

DATA SUMMARY

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Adjusted Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posttest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>15</td>
<td>25.00</td>
<td>4.56</td>
<td>29.53</td>
<td>4.86</td>
<td>28.57</td>
</tr>
<tr>
<td>Treatment</td>
<td>14</td>
<td>22.57</td>
<td>6.31</td>
<td>30.93</td>
<td>7.67</td>
<td>31.96</td>
</tr>
</tbody>
</table>

F Statistic for homogeneity of within-class regression = .845.

---

ANALYSIS OF COVARIANCE

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>78.99</td>
<td>1</td>
<td>78.99</td>
<td>3.46*</td>
</tr>
<tr>
<td>Error</td>
<td>594.29</td>
<td>26</td>
<td>22.86</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>673.28</td>
<td>27</td>
<td>24.94</td>
<td></td>
</tr>
</tbody>
</table>

*p = .074
gain scores, $t(357) = 2.20, p < .05$, and $t(354) = 2.12, p < .05$, respectively. Caution should be observed in making interpretations of these results, however, given the nature of a two-year follow-up compared to a 3.5 month follow-up.
CHAPTER V

DISCUSSION AND RECOMMENDATIONS

This chapter discusses (a) the results of the application of criterion-referenced procedures to the construction and field tryout of a diagnostic-prescriptive system for developing measurement competency for prospective teachers; and (b) possible resolutions of some technical issues in criterion-referenced measurement.

Diagnostic-Prescriptive System

One question that could legitimately be asked is whether the diagnostic-prescriptive system exemplifies resolution of the topic issues in criterion-referenced measurement or whether it further confounds the problem. To help answer this question, a description and evaluation of the process is in order.

First, behavioral objectives. The diagnostic-prescriptive system utilized as objectives the 70-statement Checklist of Measurement Competencies developed by Mayo (1967). This Checklist had been reviewed by a national sample of teachers, principals and superintendents, college professors, measurement specialists, and others. In addition, the content of several current textbooks in tests and measurements was investigated by the present writer to verify the appropriateness of using this Checklist as the basis for the learning objectives of the system. It is possible that
another investigator attempting to construct a diagnostic-prescriptive system in measurement might come up with a different, though similar, set of objectives; such is the nature of writing and selecting objectives. Nevertheless, the technique used in the present study—a judgmental approach with verification by a panel of experts—certainly represents the state of the art and resulted in a credible and defensible set of objectives.

Second, test construction. The diagnostic-prescriptive system utilized an item pool that was described in Chapter III. No manner of item analysis was performed on the diagnostic test items because the field try-out allowed the students to self-correct their work, thus providing no indication of whether the students got the items correct from knowledge or from the Answer Key. Because most of the items had been used by other sources, however, they were presumed to be "good" items. Though more rigorous procedures could have been used, expert judgment of the extent to which the items are presumed to measure the selected objectives has been considered by some investigators in the field as an acceptable method for selecting items.

Third, reliability. No reliability studies were undertaken. The purpose of the diagnostic-prescriptive system was not to classify students into "masters" versus "nonmasters"; i.e., the system was not designed to be a mastery test. The system was designed to diagnose weaknesses, which involves less rigorous assumptions than how consistent are replicated methods of placing students into particular mastery states.
Fourth, validity. Content validity was assumed, based on the manner of selecting items to measure the identified learning objectives.

Fifth, mastery criterion. Because the emphasis on the diagnostic-prescriptive system was on diagnosis, not on mastery, it was irrelevant to set minimum numbers of items correct to assume mastery. In the diagnosis approach, even one incorrect item for a particular learning objective may be indicative of a skill deficiency.

So, does the diagnostic-prescriptive system exemplify resolution of the topic issues in criterion-referenced measurement or does it further confound the problem? The opinion of this writer is that the system does neither; it represents, for better or for worse, the state of the art as we know it today.

In general, the field tryout confirmed the utility and classroom manageability of the diagnostic-prescriptive system for developing measurement competency for prospective teachers. The system was used during the entire semester without any difficulties experienced by either the students or the instructor. Prior to each examination, the students administered to themselves the diagnostic tests, scored them by using the Answer Keys, and referred to the Prescription Sheets in order to ascertain specifically which learning objective(s) they were weak in. The Prescription Sheets also provided prescriptions for remediation--page references in their textbook where the material was covered.

It was originally felt that the major value of the field tryout lay in the
confirmation of the classroom management aspects of the system. That the above results were obtained seems to justify the usefulness of the system. It should be further noted that the present study utilized a model of diagnosis-prescription-remediation often used in the elementary school situation and applied it successfully to the university level.

As an ancillary study, it was hypothesized that there would be a significant improvement in scores on the Measurement Competency Test after using the diagnostic-prescriptive system. A t-test confirmed this hypothesis ($p < .005$).

It was also hypothesized that there would be a greater increase in scores on the Measurement Competency Test for a class using the diagnostic-prescriptive system over one which did not. Here, while the results of the study were in the hypothesized direction, the difference did not achieve a significant value.

Limitations of this study should be noted, however. In this study, as in many other studies in educational research, students could not easily be randomly assigned to the treatment versus control group. In order to achieve a comparison, two different classes from two different semesters had to be utilized in a Nonequivalent Control Group Design (Campbell & Stanley, 1963). While analysis of covariance was employed in order to "equate" the two groups based on pretest scores, this technique is not the same as random assignment.

Sources of invalidity for such a design include selection-maturation
interaction and testing-treatment interaction. Other possible sources of
invalidity include regression, selection-treatment interaction, and reactive
effects of experimental arrangements. The effect of history, ordinarily
controlled in this design, may also be a source of invalidity due to the fact
that the two classes were tested in two different semesters. (See Campbell
& Stanley, 1963.)

Other limitations of the study included small Ns and the fact that
there was no extrinsic motivation for the students to do well on the Measurement Competency Test; it did not count toward the course grade. Nor
did participating in the diagnostic-prescriptive system count toward the
course grade, though students were told that some of the items on the diag­
nostic tests would appear on the final examination. Examination of the stu­
dent materials revealed that, despite the probable lack of motivation to
participate, the students did participate. That the results were in the
hypothesized direction should be viewed as encouraging, given that the
Measurement Competency Test had no bearing on course grades. And the
fact that the students and the instructor used the system with no difficulty
and seemed to find it helpful should be viewed as the most encouraging
finding of all--especially since that determination, and not the conducting
of an experiment, was the major intent of the field tryout.

Because of the encouraging nature of the field tryout, future efforts
might be directed toward providing Prescription Sheets for other widely­
used textbooks in tests and measurements. And, with the cooperation of
other graduate schools of education, a future field tryout might well be replicated on a more national basis, not only to examine the potential for improved acquisition of measurement competency for prospective teachers, but also to help determine whether a wider distribution of the materials ought to be considered.

Toward a Resolution of Some Technical Issues

Chapters II and III presented an historical background to criterion-referenced measurement and a discussion of some important psychometric issues currently confronting the field. The issues included: (a) behavioral objectives; (b) test construction; (c) reliability; (d) validity; (e) mastery criterion; and (f) norm-referenced criterion-referenced measures. The following discussion summarizes some conclusions in these areas and presents, where appropriate, recommendations for resolution. As briefly mentioned in Chapter III, there appear to be a number of parallels between criterion-referenced measurement and personnel testing; these parallels will be further explored in the discussion which follows.

Behavioral Objectives

The topics of writing objectives and selecting objectives were discussed in Chapter III. It was noted that more is known about the writing of objectives than the selecting of objectives for use in a criterion-referenced test. For the former, suggestions, referencing Mager's (1962) work, were presented. For the latter, several investigators' criteria for selecting
objectives were presented. Ultimately, the criteria for selecting objectives have to be judgments based on needs assessment. One must consider what should be mastered and what can be mastered in a particular learning situation (Guzaitis, 1973). But there is no known formula for the selection process, though curriculum guides and scope and sequence charts often provide a fertile starting point. It is highly recommended that the objectives selected be submitted to a panel of experts—subject-matter and test specialists—in order to obtain confirmation.

The parallels in personnel testing are striking. The problem of job analysis seems highly relevant to the problem of selecting behavioral objectives. In personnel selection, the first step in test validation research is the job analysis. Job analysis is a "procedure used to help identify the critical behavioral functions that make up successful performance on the job" (Shub, 1970, p. 1). Job analysis consists of (a) defining the job, and (b) discovering what the job calls for in employee behaviors and under what circumstances they are carried out. Two major purposes of job analysis are to provide insights for (a) choosing potentially useful predictors of job success, and (b) developing the criterion—measures of job performance—that one wants to predict. Job analysis may be considered equivalent to knowing the job. Just as one wants to choose behavior objectives that adequately describe the performance that one wants to teach to and test for in a criterion-referenced test, one uses a job analysis to identify those behaviors that one wants job applicants to possess or have
The job analysis procedure has been described as follows:

The job analysis involves the translation of the job description into behavioral terms—the attribute the worker needs to perform the various tasks and activities described. Through observation, interviews with workers and supervisors, and a review of pertinent manuals and literature, the job analyst seeks answers to such questions as the following:

What skills and knowledges must the new worker have at the outset? What skills and knowledges is he expected to gain during training? What physical and perceptual attributes are required by the job tasks? What mental abilities and aptitudes are needed to successfully complete training? What mental abilities and aptitudes are needed in performing the work? What personal attributes (personality, attitudes) must the employee have? (SRA, 1973, p. 4).

The analogy to criterion-referenced measurement is evident and, as in personnel testing, a great deal of professional judgment is necessary to choose the critical behaviors that one wants to measure. While personnel psychology presents guidelines and suggestions for performing the job analysis in industry (cf. Tiffin & McCormick, 1965; Lawshe & Balma, 1966; R. L. Thorndike, 1949; Dunnette, 1966; Albright, Glennon, & Smith, 1963), ultimately it is the judgment of the job analyst that determines the job elements. So it must be in educational testing.

Test Construction

The topics of item writing, number of items per objective, and item analysis were discussed in Chapter III. Generally, just as much is known about the writing of behavioral objectives, much is known about the writing
of test items. The procedures do not differ markedly from norm-referenced tests to criterion-referenced tests. In fact, because of the specificity of objectives for criterion-referenced tests, the items may well be easier to construct for a criterion-referenced test. For example, the objective "The learner will add two 2-digit numbers, sum less than five" very well defines the domain of items that measure it. Selection of items for the final edition of the test is somewhat different, however, for a criterion-referenced test than for a norm-referenced test. Some of the items that would ordinarily be thrown out by a norm-referenced-style item analysis are the very items that ought to be retained in a criterion-referenced test. Specifically, items that all or most students get correct or which all or most get incorrect would ordinarily be thrown out of a norm-referenced test as being too easy or too hard, respectively, in a test whose goal is to spread the students along an achievement continuum; because a criterion-referenced test has a goal to ascertain whether students know or do not know particular learning objectives, it is perfectly reasonable to expect that all students might get an item correct following instruction or wrong preceding instruction on that skill. Because of this situation, the traditional norm-referenced item analysis seems to have less useful relevance to criterion-referenced measurement. Instead, a sensitivity to instruction criterion, often expressed as a fourfold table, was advocated as one basis for including test items in a criterion-referenced test. As to number of items per objective, no clear-cut answer exists today; however, a rationale was presented for using three
four-choice multiple-choice items to measure each objective.

The parallel in personnel testing to test construction appears to be the construction of job sample tests. Tiffin and McCormick (1965, p. 208) defined a job sample test as

... an achievement test that consists of trying out the individual in a test situation that reproduces all, or an important sample of, the actual operations that the job itself requires. ... A scoring procedure is developed, and norms of experienced and inexperienced workers are usually obtained in the test situation as a basis of evaluating scores of persons taking the test, such as applicants.

Examples of job samples are typing tests, in-basket tests, and other job specific tests. One might view the analogy to criterion-referenced measurement in that the goal of a job sample is to distinguish between those who are experienced from those who are inexperienced; similarly, a criterion-referenced test is designed to distinguish between those who have mastered a skill from those who have not.

Reliability

The topic of assessing reliability of a criterion-referenced test was discussed in Chapter III. Various views were explored with the Harris (1974b) definition that criterion-referenced reliability is the consistency in sorting students into groups and the similar Swaminathan, Hambleton, and Algina (1974) definition that criterion-referenced reliability is consistency of placement showing the most promise.

The parallel in personnel testing might well be the reliability of employee placement and would answer the question, "How consistent are
replicated methods of placing employees into particular positions?"

Validity

The topic of assessing validity of a criterion-referenced test was discussed in Chapter III. A most important measure of the validity of a criterion-referenced test was said to be its content validity, usually ascertained by judgments of experts. Harris (1974b) referred to criterion-referenced validity as how well the test sorts students into the correct two groups (i.e., masters and nonmasters). Often a fourfold table is used, with the test decision on one axis and the "true" state on the other axis; the "true" state might be defined by teachers' judgment of students' mastery. Sensitivity to instruction was also suggested as a possible criterion-referenced validity measure, just as it was suggested as a possible criterion-referenced item analysis device.

As mentioned in Chapter III, the personnel testing's concept of synthetic validity seems related to that of content validity for criterion-referenced measures. Lawshe and Balma (1966) noted that synthetic validity involves

\[ \ldots \text{deducing a job's component parts (elements) and then inducing elemental validities into a whole.} \ldots \] Synthetic validity \ldots is capable of being viewed in a theoretical framework and of being developed through rigorous empirical investigation. The results, therefore, can be consistent, measurable, and usable. (p. 252)

Mastery Criterion

Various models of mastery were presented in Chapter III. It was
noted that the common notion of assuming mastery if 80% of the items are answered correctly is purely arbitrary. A rationale was presented for accepting three-out-of-three items correct to assume mastery if four-choice multiple-choice items are used, as suggested in the Test Construction section of Chapter III.

A parallel in personnel testing is the very important criterion development stage—the development of criteria for successful job performance. Often, a thorough job analysis will provide a list of specific, necessary job behaviors which can be used as criteria (Shub, 1970).

But why reduce the continuum of learning to only two states—mastery and nonmastery? To do so seems analogous to the personnel test's cut-off scores which yield just two categories of employees—successful and unsuccessful. There seems to be great potential in borrowing the concept of expectancy charts from personnel testing (cf. Guion, 1965). The expectancy chart is

... a practical guide for making personnel selection decisions based on the applicant's test score. An expectancy chart gives, for each test score, the probability (or chances in 100) of successful performance in the job for which the applicant is applying. (Shub, 1970, p. 3)

Analogously, one could set up a criterion-referenced research program to yield statements such as "John got two out of three items correct dealing with addition of two 2-digit numbers; the probability that he has mastered this skill is 0.72." Or, "Marcia got three out of three items correct dealing with addition of two 2-digit numbers; the probability that she has mastered this skill is 0.95." And so on. Thus, instead of being forced
to create a mastery-nonmastery decision, one can represent "mastery" as it really is in nature--a continuum of knowledge.

**Norm-Referenced Criterion-Referenced Measures**

The movement toward norm-referenced criterion-referenced measures was discussed in Chapter III. It is probably inevitable that in the not so distant future we will see criterion-referenced test scores accompanied by an indication of "how good" the performance is--i.e., characterized by norms. In industry, one might look not only at the job applicant's probability of success on the job but also how high his test score placed him within the population of job applicants--though admittedly the latter is probably not as important in industry as in education.

**Concluding Note**

The goal of this dissertation was to present and analyze the relevant technical issues in criterion-referenced measurement with a view toward recommending some resolution and applying the state of the art to producing a diagnostic-prescriptive system for developing measurement competency for prospective teachers. One recurring finding throughout this dissertation should be that there are no final answers yet; this dissertation attempted to point the way toward resolution, with an eye toward borrowing some concepts from personnel testing. The problems of criterion-referenced measurement are not unlike the problems of this much older discipline. This is not to say that personnel psychologists have solved the
major problems. Job analysis and criterion development are still very much an art and may well continue to be so; similarly, many of the concepts of criterion-referenced measurement may well remain an art as opposed to a science. Perhaps some sort of joint effort between professionals of these two disciplines will result in a major breakthrough in both these areas. What is ultimately needed is a criterion-referenced version of Gulliksen's (1950) *Theory of Mental Tests*. To accomplish this, further research and improved technology are required.
SUMMARY

Accountability. Guaranteed performance. Pay for performance. Relevance. Individualized instruction. Test scores. The issues reflected by these terms have become burning ones in recent years as parents and educators alike are demanding that the educational process be increasingly responsive to the individual needs of children. How do we determine these needs? The criterion-referenced measurement point of view suggests one solution.

From all the recent discussions in journals and with educational test publishers now emphasizing criterion-referenced tests, one would think they are something new. Actually, they have been around for a long time—as far back as the 23rd Century B.C.—although not always by that name. Though an old concept, criterion-referenced measurement is a relatively young science and has its unique measurement problems. The purpose of this dissertation was to identify, review, and discuss the major psychometric issues and controversies in criterion-referenced measurement with a view toward recommending some resolutions. A secondary purpose was to apply the state of the art to producing a diagnostic-prescriptive system for developing measurement competency for prospective teachers.

The discussion of important technical issues currently confronting
the criterion-referenced measurement field included: (a) behavioral objectives, including writing and selecting objectives; (b) test construction including item writing, number of items per objective, and item analysis; (c) reliability; (d) validity; (e) mastery criterion; and (f) norm-referenced criterion-referenced measures. In addition to presenting some recommendations for resolution of these issues, a number of parallels between the problems of criterion-referenced measurement and those of personnel testing were noted and discussed.

A diagnostic-prescriptive system for developing measurement competency for prospective teachers was constructed. The system was designed for use in undergraduate tests and measurements courses. It consists of diagnostic tests covering 70 major learning objectives, answer keys, prescription sheets, and an instructor's guide. In general, a field tryout confirmed the utility and classroom manageability of the system. Using pre-post survey test measures, students showed significant gains after using the system. Compared to a control group, the gains were in the hypothesized direction, though not statistically significant.
REFERENCES

Adkins, D. C. Test construction: Development and interpretation of achievement tests (2nd ed.). Columbus, Ohio: Charles E. Merrill, 1974.


Fremer, J. Criterion-referencing. CSEA Newsletter, 1972, 1(2).


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

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


Larsen, V. S. Personal communication, 1971. (a)

Larsen, V. S. Personal communication, November 1971. (b)


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. (b)

Livingston, S. A. Reply to Shavelson, Block, and Ravitch's "Criterion-referenced testing: Comments on reliability." Journal of Educational Measurement, 1972, 9, 139-140. (c)


APPENDIX A
CHECKLIST OF MEASUREMENT COMPETENCIES

Directions:
Please respond to the statements below in terms of the knowledge, ability, and understanding which you believe the beginning teacher with a Bachelor's degree should possess.

Using an "X" mark, indicate whether you believe that each of the competencies "Is Essential," "Is Desirable," or "Is of Little Importance" to the work of the beginning teacher. If you do not understand some part of the statement check with an "X" in the last column at right entitled "Do Not Understand Statement." Also circle the part or parts of the statement which you do not understand. You may also wish to qualify your responses by writing in comments. If you wish to add any competencies which should have been included, feel free to do so on separate pages.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Is Essential</th>
<th>Is Desirable</th>
<th>Is of Little Importance</th>
<th>Do Not Understand Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Knowledge of advantages and disadvantages of standardized tests.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Ability to compare standardized with teacher-made tests and choose appropriately in a local situation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Ability to interpret achievement test scores.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Understanding of the importance of adhering strictly to the directions and stated time limits of standardized tests.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Knowledge of sources of information about standardized tests.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Knowledge of general information about group intelligence tests.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Knowledge of general information about individual intelligence and aptitude tests.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Familiarity with need for and application of personality and interest inventories.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Familiarity with need for and application of projective techniques.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Knowledge of general uses of tests, such as motivating, emphasizing important teaching objectives in the minds of pupils, providing practice in skill, and guiding learning.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Knowledge of advantages and disadvantages of teacher-made tests.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Knowledge of the fact that test items should be constructed in terms of both content and behavior.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Ability to state measurable educational objectives.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Knowledge of the general principles of test construction (e.g., planning the test, preparing the test and evaluating the test).</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Knowledge of advantages and disadvantages of various types of objective test items.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Knowledge of the techniques of administering a test.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17. Ability to construct different types of test items.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Is Desirable Is of Little Importance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>-----------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td>Understanding and application of correction-for-guessing formula to an objective test.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19.</td>
<td>Knowledge of the principles involved in scoring subjective and objective tests.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20.</td>
<td>Knowledge of effective procedures in reporting to parents.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22.</td>
<td>Knowledge of advantages and disadvantages of essay questions.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23.</td>
<td>Familiarity with the blueprint scheme for dealing with the content and behavior dimensions in test planning.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.</td>
<td>Ability to interpret diagnostic test results so as to evaluate pupil progress.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.</td>
<td>Ability to interpret the ratio formula relating CA, MA and IQ.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.</td>
<td>Familiarity with expected academic behavior of students classified in certain IQ ranges.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27.</td>
<td>Ability to interpret a profile of sub-test results of standardized tests.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28.</td>
<td>Knowledge of limitations of tests that require reading comprehension.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29.</td>
<td>Understanding of the limitations of the &quot;percentage&quot; system of marking.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30.</td>
<td>Understanding of the limitations of applying national norms to a local situation.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31.</td>
<td>Ability to compare two classes on the basis of the means and standard deviations of a test.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32.</td>
<td>Knowledge of concepts of validity, reliability and item analysis.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33.</td>
<td>Ability to do a simple item analysis for a teacher-made test.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34.</td>
<td>Knowledge of the limitations of ability grouping based on only one measure of ability.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35.</td>
<td>Knowledge of limitations in interpreting IQ scores.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36.</td>
<td>Familiarity with the nature and uses of a frequency distribution.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37.</td>
<td>Familiarity with techniques of ranking a set of scores.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38.</td>
<td>Ability to set up class intervals for a frequency distribution.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Understanding of the basic concept of the standard error of measurement.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Understanding of the nature and uses of the histogram and frequency polygon.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Understanding of the nature and uses of the mode, median and mean.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Ability to compute the mode, median and mean for simple sets of data.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Knowledge of advantages and disadvantages of the mode, median and mean.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Understanding of the meaning of the term &quot;variability&quot; and its connection with such terms as &quot;scatter,&quot; &quot;dispersion,&quot; &quot;deviation,&quot; &quot;homogeneity&quot; and &quot;heterogeneity.&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>Understanding of the nature and uses of the semi-interquartile range.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Understanding of the nature and uses of the standard deviation.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>Ability to compute the semi-interquartile range for simple sets of data.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>Knowledge of the approximate percentile ranks associated with standard scores along the horizontal baseline of the normal curve.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>Knowledge of the percentage of the total number of cases included between ± or - 1, 2 or 3 standard deviations from the mean in a normal distribution.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>Knowledge of the fact that the normal curve is an ideal distribution, an abstract model approached but never achieved fully in practice.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>Knowledge of the limitations of using the normal curve in practice as the fact that in large heterogeneous groups it &quot;fits&quot; most test data rather well and that it aids in the interpretation of test scores, but does not necessarily apply to small selected groups.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>Ability to convert a given raw score into a z score from a mean and standard deviation of a set of scores.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td>Knowledge of the means and standard deviations of common standard score scales such as the z, T, stanine, deviation IQ and CEEB scales.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16.</td>
<td>Knowledge of the common applications of standard scores.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17.</td>
<td>Knowledge of how to convert from one type of standard score to another.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td>Knowledge of the fact that the mode, mean and median coincide for a symmetrical distribution.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
57. Knowledge of the meaning of the terms used to designate certain common non-normal distributions such as "positively skewed," "negatively skewed," and "bimodal" distributions.

58. Knowledge of the fact that any normal distribution can be completely described in terms of its mean and standard deviation.

59. Ability to define the concept of correlation, including such terms as "positive correlation," "negative correlation," "no relationship" and "perfect relationship."

60. Knowledge of the significance of the numerical magnitude and the sign of the Pearson Product-Moment Correlation Coefficient.

61. Knowledge of the fact that correlation coefficients do not imply causality between two measures.

62. Knowledge of the fact that correlation coefficients alone do not indicate any kind of percentage.

63. Understanding of the meaning of a given correlation coefficient in terms of whether it is "high," "low" or "moderate."

64. Familiarity with the scatter diagram and the ability to make simple interpretations from it.

65. Knowledge of what size of correlation to expect between two given variables in terms of logical reasoning, e.g., in terms of a common factor.

66. Understanding of the fact that a raw score has no meaning alone and needs some context in which it can be interpreted.

67. Familiarity with the nature and uses of the common derived scores, viz., age scales, grade scales, percentile scales and standard score scales.

68. Understanding of certain concepts associated with scale theory, such as types of scales (nominal, ordinal, cardinal and absolute); translation of scores to a common scale; units of equal size; and common reference points (zero or the mean.)

69. Ability to interpret raw scores from a given set of norms.

70. Understanding of the fact that interpretation of achievement from norms is affected by ability level, cultural background and curricular factors.
APPENDIX B
MEASUREMENT COMPETENCY TEST - Form A

DIRECTIONS: Make no marks on this test booklet. Print your name, the title of the examination MEASUREMENT COMPETENCY TEST FORM A, your college and city in the margin of the IBM answer sheet.

Two different kinds of objective test items comprise this test. They are the multiple-choice and the key-list types. It is essential that you follow the directions carefully as you go from a set of one type of item to a set of the other. If you do not understand the specific directions, ask the proctor for an explanation.

When marking your answers on the IBM answer sheet, you should use either an electrographic pencil, if you have one, or another type of soft black graphite pencil. Do not use a ball point pen or a wax pencil. Make your marks thus:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Solid black marks are made by going over each mark two or three times and by pressing firmly on your pencil.
2. If you change your mind, erase your first mark completely.
3. Make no unnecessary marks in or around the dotted lines. Do not rest your pencil on a lettered space while deciding which space to mark.
4. Keep your answer sheet on a hard surface while marking your answers.
5. Make your marks as long as the pair of dotted lines.
6. Make one mark and only one mark after each answer sheet number.
7. You are expected to answer every item. If you are not sure of the answer to an item, put down the answer that seems most likely to you.
8. Try to change as few of your answers as possible. Your first impressions are usually best.

Department of Education
Loyola University, Chicago
After the number on the answer sheet which corresponds to that of each of the following items, blacken one lettered space to indicate the correct answer.

1. The essential difference between standardized and unstandardized tests lies in
   A. their validity.
   B. their objectivity.
   C. the availability of norms.
   D. the discriminatory capacity of their items.

2. Advocates of "culture fair" tests of mental ability can most justifiably criticize the Stanford-Binet because of its emphasis in measuring
   A. organization of ideas.
   B. fluency of ideas.
   C. verbal abilities.
   D. innate abilities.

3. If a student wanted to find the most appropriate achievement test in arithmetic, he should consult
   A. publishers' catalogues.
   B. Buros' Mental Measurements Yearbook.
   C. Journal of Experimental Education.
   D. the most recent texts in the teaching of arithmetic.

4. If a teacher wanted to determine how well a standardized test would measure the objectives which she had been trying to teach, it would be best for her to examine
   A. the test itself.
   B. critical reviews of the test.
   C. the manual for the test.
   D. recent studies in which the test had been used.

5. The type of measuring device considered to require the most technical knowledge for its administration and interpretation is
   A. a group intelligence test.
   B. a self-report personality inventory.
   C. a projective test of personality.
   D. a survey achievement battery.

6. The distinction between aptitude and achievement tests is chiefly one of
   A. purpose for which used.
   B. type of ability measured.
   C. method of measurement.
   D. breadth of content.

7. Two general types of achievement tests have been used in secondary grades. These are (1) tests of knowledge of content common to many textbooks, and (2) tests requiring application and interpretation. What is the current status of the two types of tests?
   A. Most current tests are of type 1 and current emphasis is in the direction of type 1.
   B. Most current tests are of type 1 but current emphasis is in the direction of type 2.
   C. Most current tests are of type 2 but current emphasis is in the direction of type 1.
   D. Most current tests are of type 2 and current emphasis is in the direction of type 2.

8. High interest inventory scores relevant to a given occupation are most likely to be predictive of
   A. success in training for the occupation.
   B. actual future employment in the specific occupation.
   C. degree of success within the occupation.
   D. satisfaction with the occupation, assuming employment and requisite ability.

9. Scores on standardized intelligence tests are based on the assumption that all pupils
   A. have had some experience with such tests.
   B. have had some formal schooling.
   C. have had similar backgrounds of experience.
   D. are unfamiliar with the test material.
10. Which one of the following scores appearing in a student's record would be most meaningful without further reference to the group?
   A. 23 items correct in an English test of 40 items.
   B. 30 items wrong in an algebra test of 50 items.
   C. 100 words per minute in a typing test.
   D. Omitted ten items in each of the English and algebra tests.

11. The Navy reports aptitude test results in terms of standard scores with a mean of 50 and a standard deviation of 10. A recruit with mechanical comprehension score of 65 is a candidate for machinist training. On the basis of this score he would be judged
   A. a very promising candidate.
   B. slightly above average.
   C. average.
   D. slightly below average.

For each of the following paired items, blacken one lettered space to indicate that the first item is
   A. greater than the second
   B. less than the second
   C. definitely equal to the second
   D. of uncertain size with reference to the second

12. Usefulness of survey achievement batteries in providing data useful in guidance on the high school level.
    Usefulness of survey achievement batteries in providing data useful in assigning grades on the high school level.

13. The amount of structuring in a non-projective personality test.
    The amount of structuring in a typical projective personality test.

14. Usefulness of a vocational interest inventory in predicting vocational success.
    Usefulness of a vocational aptitude test in predicting vocational success.

15. Importance of the physical conditions of the room upon test performance.
    Importance of health factors upon test performance.

After the number on the answer sheet which corresponds to that of each of the following items, blacken one lettered space to indicate the correct answer.

16. It is more appropriate to discuss the mental stanine of a child with a parent than the child's I.Q. because
   A. the stanine is a more valid measure of intelligence.
   B. the I.Q. appears more precise than it actually is.
   C. mental stanines are more highly correlated with achievement.
   D. parents are better kept in doubt with reference to the child's ability.

17. What is the major argument for using unstructured essay exercises in tests given during instruction?
   A. Unstructured exercises insure that students attack the same problems.
   B. Teacher insights with reference to student thought patterns and attitudes are promoted.
   C. Course marks are more valid measures of student ability.
   D. Such exercises best stimulate students to write well-organized essay answers.

18. Why is it most desirable to use such words as "contrast," "compare" and "criticize" in formulating essay exercises?
   A. Such words are readily understood by students.
   B. Such words tend to characterize unstructured exercises.
   C. Such words stimulate students to recall relevant facts.
   D. Such words tend to characterize thought rather than fact questions.
19. How reliably can answers to essay questions be evaluated?
   A. It is impossible to evaluate them reliably enough to justify the use of this form.
   B. Under certain conditions they can be evaluated reliably, but the process is likely to be
difficult and costly.
   C. They can be evaluated reliably with great ease if certain simple precautions are observed.
   D. They are ordinarily evaluated with as much reliability as are objective tests.

20. Which of the following types of items is well adapted to evaluating student knowledge of
    numerous technical terms?
   A. True-false.
   B. Multiple-choice.
   C. Matching.
   D. Analogies.

21. The term objective, when used to label an educational test, describes
    A. a characteristic of the scoring process.
    B. a typographic feature of the test.
    C. the degree of standardization of the test.
    D. the content limitations of the questions.

22. Sue answered correctly 25 out of 50 items on an arithmetic test. What interpretation can be
    made of Sue's performance on the test?
    A. Sue placed at the 50th percentile.
    B. Sue needs remedial work in arithmetic.
    C. Sue knows about one-half of the material in arithmetic taught in her grade.
    D. No interpretation of the score is possible on the basis of the information given.

23. Which of the following is a poor suggestion for the construction and use of essay examinations?
    A. Restrict the use of the essay examination to those levels of knowledge to which it is best
    adapted.
    B. Make definite provisions for teaching pupils how to take examinations.
    C. Increase the number of questions asked but restrict the possible answers.
    D. All of these are good suggestions.

24. Problems arise in attempting to develop measures of ultimate goals mainly because
    A. measurement methods have not given proper weight to all goals.
    B. teachers have been reluctant to depart from traditional testing methods.
    C. group norms with which to compare results are not available.
    D. such goals concern behavior not usually observable under classroom conditions.

25. Which of the following is an untrue statement about instructional goals?
    A. The worth of a goal is determined by its measurability.
    B. A two-way chart helps to relate content to educational goals.
    C. One test can usually measure only a few goals.
    D. Content and method vary directly with goals.

26. Why should behavioral objectives as contrasted with content objectives best be restricted in
    number?
    A. To facilitate organization of a course.
    B. To promote their operational definition.
    C. To enable a teacher to keep them constantly in mind during instruction.
    D. There are few basic factors in human ability.

27. "Washington, D.C., is the most important city in the United States." Why is this a poor
    true-false item?
    A. It is ambiguous.
    B. It is too easy.
    C. It is too brief.
    D. It is too factual.

28. "Philadelphia was the capital and largest city in the United States for a number of years." Why
    is this a poor true-false item?
    A. It is ambiguous.
    B. It involves more than one idea.
    C. It does not have a good answer.
    D. It is too long.
29. "The capital of New York State is
1. Albany.
2. Buffalo.
3. Chicago.
4. New York City."
What would be the best change to make in this item?
A. Add the word "at" to the stem.
B. Rewrite stem to read "Which city is the capital of New York State?"
C. Replace "Chicago" with "Rochester."
D. Replace "New York City" with "Syracuse."

30. "In the United States, ______ are elected for ______ and ______ for ______."
What would be the best way to revise this item?
A. Replace the first blank by "senators" and the third blank by "representatives."
B. Insert the word "years" after the second and fourth blanks.
C. Insert the word "all" before the first and third blanks.
D. Make changes A and B.

31. Validity is determined by finding the correlation between scores on
A. the even numbered items on a test and the odd numbered items on that test.
B. one form of a test and another form of that same test.
C. a test and some independent criterion.
D. two administrations of the same test.

32. What is most wrong with the statement, "This test is valid."
A. The statement does not specify what the test is valid for.
B. The word "valid" is vague. A numerical coefficient should be given.
C. A test does not show validity or lack of it.
D. The statement is meaningless, since it does not specify the conditions of administration.

33. For determining reliability, for retesting doubtful cases, or for measuring growth, it is
most useful to have
A. equivalent forms.
B. adequate norms.
C. objectivity and interpretability.
D. logical and empirical validity.

34. If the reliability of an arithmetic test is .50, and if the length is doubled, the reliability
would
A. increase.
B. decrease.
C. remain the same.
D. change in some indeterminate way.

35. A spelling test is given twice within a few days to a third-grade pupil. The first time he
receives a second-grade rating. His second performance puts him at the fourth-grade level.
The test is probably
A. unreliable.
B. lacking in validity.
C. not objective.
D. one easily remembered.

36. Upon receiving intelligence test scores for her class a teacher is surprised to learn that a
pupil she has always considered as "average" has an I.Q. of 84. or the following, what is
her most appropriate course of action?
A. Check the pupil's cumulative record for the results of previously administered achievement
and intelligence tests.
B. Evaluate her attitude toward the pupil's performance in class to learn whether she has
been grading him too leniently.
C. Discuss the test results with the pupil to learn whether he was ill on the day of the test.
D. Recognize that the pupil is achieving far beyond his capacity and encourage him to continue.
37. What is the chief obstacle to effective homogeneous grouping of pupils on the basis of their educational ability?
A. Resistance of children and parents to discriminations on the basis of ability.
B. Difficulty of developing suitably different teaching techniques for the various levels.
C. Increased costs of instruction as the number of groups increases and their average size decreases.
D. Wide differences in the level of development of various abilities within individual pupils.

38. A diagnostic test which provides the teacher with a profile of scores is of little value unless:
A. the sub-tests which make up the profile are quite reliable.
B. the test has reliable norms.
C. the test has been shown to be a valid predictor of future achievement.
D. the scores are reported in terms of percentile ranks.

39. Peter is exactly 10 years old. His mental age is 12 years 6 months. What is his ratio I.Q.?
A. 80
B. 95
C. 125
D. None of the above.

40. In order to compute a correlation coefficient between traits A and B, it is necessary to have:
A. measures of trait A on the group of persons, and of trait B on another.
B. one group of persons, some who have both A and B, some with neither, and some with one but not the other.
C. two groups of persons, one which could be classified as A or not A, the other as B or not B.
D. measures of traits A and B on each person in one group.

41. Test norms are most satisfactory when the sample of pupils or students used in establishing the norms:
A. consists of nearly all pupils or students taking the test prior to the time the norms are published.
B. is representative of a clearly defined population with which it is appropriate to make comparisons.
C. ranges over all the grade levels in which the test is likely to be used.
D. includes all schools volunteering to participate in the standardization testing.

42. A good diagnostic test differs from a good survey achievement test in:
A. reliable and valid measurement of skills.
B. identifying causes of weaknesses.
C. possessing equivalent forms so that growth in achievement can be measured.
D. identifying pupils whose achievement is unsatisfactory.

43. Item difficulty values (percents of correct responses to each test item) are useful in:
A. evaluating attainment of instructional objectives.
B. arranging items in order of difficulty.
C. revising a series of items.
D. accomplishing all of the above.

44. On a given test item, 30 per cent of the top fourth of the pupils marked the correct answer, and 70 per cent of the lowest fourth responded correctly. The discriminating power of the item is:
A. decidedly negative.
B. slightly negative.
C. definitely positive.
D. almost perfect.

45. The State of X has a state-wide testing program. As a basis for revising the objective examination in science, a set of papers from the top and bottom quarter of the total group tested was analyzed. The per cent passing each item was determined. Other things being equal, which of the following items would one be most likely to keep in the test?
A. Top quarter -- 95%, bottom quarter -- 92%
B. Top quarter -- 80%, bottom quarter -- 60%
C. Top quarter -- 70%, bottom quarter -- 75%
D. Top quarter -- 25%, bottom quarter -- 10%
For each of the following items, blacken one lettered space to indicate that the item correctly refers to

A the mean
B the median
C the standard deviation
D the quartile deviation
E more than one of the above

Be sure to consider the possibility that "E" is the correct answer.

46. Is the point on the scale of measurement above which and below which there are fifty per cent of the cases.

47. An example of a measure of "central tendency."

48. Is especially useful as an average where a distribution of test scores includes a number of extremely high scores or extremely low ones.

49. Can be used in comparing their performance on a test of mental ability if computed for two different groups.

50. When computed from a frequency distribution, it is necessary at one stage to multiply by the number of units in a class interval.

51. Is represented by a distance of 10 T-score units, 2 stanine units and one z-score unit.

After each exercise number on the answer sheet, blacken one lettered space to designate the correct answer.

52. In the set of scores: 27, 50, 13, 5, 34, 63, the median is closest to
   A. 29
   B. 34
   C. 35.1
   D. 36.5

53. Scores on standardized tests used in the elementary schools are most often converted to grade scores, for example, 6.6 or 7.3 rather than to percentile ranks. On the high school level the scores are usually converted to percentile ranks. Why?
   A. Differences in percentile ranks are in terms of equal units of ability.
   B. Grade scores assume common educational experience over the years; percentile ranks do not.
   C. Percentile ranks are necessarily more reliable than grade scores.
   D. Percentile ranks can more easily be converted to percent marks.

54. Which of the following types of derived measures is least used at the present time?
   A. Achievement quotient.
   B. Grade score.
   C. Intelligence quotient.
   D. Scaled score.

55. Find the mean of a grouped frequency distribution if the interval is 5, the arbitrary origin was taken at 25, the sum of the deviations about the arbitrary origin is 10 and the number of cases is 50.
   A. 24
   B. 25
   C. 26
   D. 27
56. A student scores 35 on a vocabulary test. The mean for the class is 37.3 and the standard deviation is 8.4. His z-score is
   A. .27  
   B. .23  
   C. -.27  
   D. -.18

57. What does the percentile equivalent of a raw score indicate?
   A. The per cent of a group making scores above the mid-point of that raw score interval. 
   B. The per cent of a group making scores between the upper and lower limits of that raw score interval. 
   C. The per cent of a group making scores lower than the mid-point of that raw score interval. 
   D. The per cent of items of the test which must be answered correctly to get that raw score.

58. In a particular situation the frequency distribution of scores on a standardized test is found to be approximately normal. This should be regarded as
   A. common and highly desirable. 
   B. common but not especially desirable. 
   C. rare and highly desirable. 
   D. rare and not especially desirable.

59. If a certain test is taken by a group of high school seniors, and is found to correlate .82 with freshman grades received in college by these same seniors, one can say that
   A. the test is a valid predictor of college aptitude. 
   B. the test is not a reliable measure of college success. 
   C. approximately two-thirds of those taking the test will be successful in college. 
   D. students who score lower than 82 will be unsuccessful in college.

60. The standard error of measurement is a numerical figure which indicates
   A. the number of points a student's test score is in error in relation to the score he should make. 
   B. the number of points the mean score for the test is in error. 
   C. a range of scores within which the student's true score most probably falls. 
   D. the reliability of the test norms.

---

When you have finished the test, make sure that each mark on your answer sheet is solid, black, and glossy. Erase all superfluous marks on the answer sheet no matter how tiny. Then turn in your booklet and answer sheet to the proctor in charge.
MEASUREMENT COMPETENCY TEST - Form B

DIRECTIONS: Make no marks on this test booklet. Print your name, the title of the examination MEASUREMENT COMPETENCY TEST FORM B, your college and city in the margin of the IBM answer sheet.

Two different kinds of objective test items comprise this test. They are the multiple-choice and the key-list types. It is essential that you follow the directions carefully as you go from a set of one type of item to a set of the other. If you do not understand the specific directions, ask the proctor for an explanation.

When marking your answers on the IBM answer sheet, you should use either an electrographic pencil, if you have one, or another type of soft, black graphite pencil. Do not use a ball point pen or a wax pencil. Make your marks thus:

A B C D E

1. Solid black marks are made by going over each mark two or three times and by pressing firmly on your pencil.

2. If you change your mind, erase your first mark completely.

3. Make no unnecessary marks in or around the dotted lines. Do not rest your pencil on a lettered space while deciding which space to mark.

4. Keep your answer sheet on a hard surface while marking your answers.

5. Make your marks as long as the pair of dotted lines.

6. Make one mark and only one mark after each answer sheet number.

7. You are expected to answer every item. If you are not sure of the answer to an item, put down the answer that seems most likely to you.

8. Try to change as few of your answers as possible. Your first impressions are usually best.

Department of Education
Loyola University, Chicago
After each exercise number on the answer sheet, blacken one lettered space to designate the correct answer.

1. Which of the following types of norms is least effective on the high school level?
   A. Percentile ranks.
   B. Stanines.
   C. T-scores.
   D. Grade scores.

2. The standard deviation of I.Q.'s on the Binet scale of a representative sample of white urban school children has been found to be about 16. This means that approximately 34% of the cases will have I.Q.'s between
   A. 92 and 108
   B. 84 and 116
   C. 84 and 100
   D. 100 and 132

3. A graphical device showing the distribution of scores on a single test is called a
   A. scattergram.
   B. histogram.
   C. line graph.
   D. frequency table.

4. Under a scattergram there is a notation that the coefficient of correlation is .06. This means that
   A. most of the cases are plotted within a range of 6% above or below a sloping line in the diagram.
   B. plus and minus 6% from the means includes about 68% of the cases.
   C. there is a negligible correlation between the two variables.
   D. most of the data plotted fall into a narrow band 6% wide.

5. A teacher is in the habit of giving his geometry students a weekly test. In the middle of the school year, six of the students in his class transfer to another school. For the remaining students, which of the following will probably show the greatest amount of change?
   A. The raw score they make on the weekly tests.
   B. Their rank in class as determined by the weekly tests.
   C. The average weekly test scores.
   D. The range of their weekly test scores.

6. In a frequency distribution representing a group of 50 individuals, the median is in the score interval whose indicated limits are 45-52. The number of cases up to the lower limit of this interval is 18, and there are ten cases in this interval. What proportion of the 45-52 interval falls below the median?
   A. 30%
   B. 50%
   C. 70%
   D. Indeterminate from the data given.

7. A student's raw score is exactly in the middle of the range of raw scores assigned a stanine of 7. If his raw score were assigned a T-score, it would be numerically equal to
   A. 30
   B. 40
   C. 60
   D. 75

8. In a frequency distribution of 250 scores, the mean is reported as 78 and the median as 65. One would expect this distribution to be
   A. positively skewed.
   B. negatively skewed.
   C. symmetrical.
   D. normal.
9. Which of the following shows the highest degree of correlation?
   A. +.40
   B. -.20
   C. -.50
   D. -.65

10. Below are the percentile scores of four students on a standardized reading test:
   Mary: 45  Tom: 90
   Jane: 50  Jim: 95

   What can be said about the difference in these students' achievement?
   A. The relative differences in achievement between Mary and Jane is equal to that between Tom and Jim.
   B. Tom's achievement is twice as great as Mary's.
   C. The teacher can be more certain about Jim being better than Tom than she can about Jane being better than Mary.
   D. The teacher should recognize that if the test were administered a second time, it is quite probable that Tom would do better than Jim.

For each of the following items, blacken one lettered space to indicate that the item correctly refers to
   A the mean
   B the median
   C the standard deviation
   D the quartile deviation
   E more than one of the above

   Be sure to consider the possibility that "E" is the correct answer.

11. Includes approximately 68 per cent of the cases when measured above and below the mean in a normal distribution.

12. May be obtained by summing the scores and dividing by the total number of scores.

13. Is most often confused with the "mid-score."

14. A point that is affected markedly by extremely high or low scores.

15. Is represented by a T-score of 50, a stanine of 5 and a z-score of 0.

After each exercise number on the answer sheet, blacken one lettered space to designate the correct answer.

16. At the end of the semester a history teacher gave his pupils an essay test on the material covered during the preceding weeks. When he graded the papers he deducted points from the total score for spelling, grammar and English usage. In so doing, he
   A. increased the accuracy of his final grades.
   B. increased the objectivity of measurement.
   C. lowered the reliability of the test.
   D. lowered the validity of the test.

17. A teacher has given four 100-item achievement tests with the following results. Which test apparently was most suitable for the group?
   A. Test I: mean, 60; range, 17-80
   B. Test II: mean, 58; range, 18-82
   C. Test III: mean, 68; range, 26-99
   D. Test IV: mean, 88; range, 62-98
18. John scored at the 60th percentile on an academic aptitude test and scored at the 57th percentile on a test of reading ability. The above data indicate that John's teacher should
A. ignore this difference altogether.
B. provide him with individual help in reading.
C. motivate him to read more extensively outside of school.
D. have him retested in reading ability.

19. The same test is given on successive days to the same class. The correlation between the two sets of scores is .95. Which conclusion concerning the scores is most defensible?
A. They are highly reliable.
B. They are highly valid.
C. They are quite unstable.
D. They are not differentiating.

20. An achievement test item is characterized by the following item analysis data where B is the keyed answer:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Group</td>
<td>8</td>
<td>17</td>
<td>19</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>Low Group</td>
<td>16</td>
<td>19</td>
<td>24</td>
<td>26</td>
<td>15</td>
</tr>
</tbody>
</table>

One can infer from the data given above, that this item
A. is a relatively easy one.
B. has distractors all needing revision.
C. is of satisfactory discriminating power.
D. has not been keyed correctly.

21. In tallying a frequency distribution of test scores, class intervals of 15-19, 20-24, 25-29, etc. are used. Where 22, rather than 22.5, is taken as the mid-point of the interval, the crucial assumption is that
A. the score of 22 means a range of 22,000 to 22,999...
B. the score 22 means a range from 21,000... to 22,000...
C. the interval 20-22 means a range from 20,000... to 22,999...
D. the interval 20-22 means a range from 19,500... to 21,499...

22. Quite often test manuals give analyses of the sources from which the items in a test have been drawn and include information with respect to the proportions of items relevant to different categories. This information is most useful in evaluating a test with respect to its
A. predictive validity.
B. content validity.
C. construct validity.
D. concurrent validity.

23. A deviation I.Q. indicates
A. deviation of MA from CA.
B. deviation of two sets of scores from the mean.
C. the distance in standard score units of a score from the mean.
D. relative achievement of a person in terms of standard score units.

24. The distributions shown differ in
A. skewness only.
B. variability only.
C. central tendency only.
D. both variability and central tendency.

25. In general, increasing the length of a test will make it more
A. valid.
B. reliable.
C. objective.
D. diagnostic.
26. A teacher is examining the manual for a new diagnostic reading test. In the section labeled, "Description of Test" she finds the statement: "This test provides measures of four completely independent reading skills." In the section labeled, "Test Statistics" she finds the following data on the reliability and intercorrelation of the four scores:

<table>
<thead>
<tr>
<th>Reading Skills</th>
<th>Par. Mean</th>
<th>Sent. Mean</th>
<th>Vocab. Mean</th>
<th>R. Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paragraph Meaning</td>
<td>.88*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Meaning</td>
<td>.80</td>
<td>.82*</td>
<td>.88*</td>
<td></td>
</tr>
<tr>
<td>Reading Vocabulary</td>
<td>.82</td>
<td>.76</td>
<td>.76</td>
<td>.94*</td>
</tr>
<tr>
<td>Reading Speed</td>
<td>.78</td>
<td>.72</td>
<td>.76</td>
<td></td>
</tr>
</tbody>
</table>

*The entries in the diagonal are reliability coefficients.

On the basis of the material in the test manual, what criticism should the teacher make?
A. The test does not measure independent reading skills.
B. The test is highly speeded.
C. The test is not sufficiently reliable to make comparisons between individual pupils.
D. The correlations among the scores indicate that the test possesses little validity.

27. Because no standardized test possesses perfect reliability it is essential that the teacher regard the score which a student obtains as
A. having little meaning unless it is very high or very low.
B. indicating a point in the range near which the student's true score probably falls.
C. indicating only that the student has either more or less ability than the average individual in the norming group.
D. providing information about the student which can be used only by a thoroughly trained guidance counselor.

28. In which of the following instances is a teacher most justified in requiring all students to make test scores of 75% or better?
A. The class is composed of above average students.
B. The questions are essay rather than objective.
C. The questions measure knowledge of essentials.
D. The pupils have ample time to prepare for the test.

29. John tells his mother that he made a score of 68 on his science test. Which type of information would best help his mother to understand the meaning of his score in terms of his achievement in science?
A. The test consisted of 90 questions.
B. Half of the class failed the test.
C. The mean score for the class was 65.
D. The highest score in the class was 83.

30. Year after year the mean achievement test scores for the students in school X consistently are one year or more above the national norms. What is the most probable cause of this finding?
A. School X is located in an upper-middle-class community.
B. School X is staffed with expert teachers.
C. School X is using tests that have unreliable norms.
D. School X stresses the traditional, rather than the activity, curriculum.

31. Which of the following is a poor principle to use in marking or assigning grades?
A. Letter grades have definite advantages over percentage grades.
B. Marks should be based as much as possible on objective measures.
C. Marks should indicate achievement of general as opposed to specific objectives.
D. Status and improvement should be graded separately.

32. Objective test exercises are most likely to measure the ability of the pupils to reason if the exercises
A. are of the recall rather than of the recognition type.
B. are similar in form to intelligence test exercises.
C. are of the multiple-answer rather than the true-false type.
D. require application of facts to a novel situation or problem.
33. The use of the normal curve as a basis for assigning school marks is most legitimate when 
   A. a standardised test is used.  
   B. all of the pupils have approximately the same I.Q.  
   C. the marks are to be assigned to a large and representative group of pupils. 
   D. the average pupil scores 85 on the test used.

34. The most important advantage of the objective test over the essay test is that it 
   A. saves time for the teacher.  
   B. has higher content validity.  
   C. measures a greater range of instructional objectives.  
   D. provides for a more complete sampling of content.

35. A two-way chart is used in identifying for each item of an achievement test the topics and 
   the behavioral objectives to which each item is relevant. The process is one of estimating 
   the test's 
   A. concurrent validity.  
   B. predictive validity.  
   C. content validity.  
   D. construct validity.

36. In the scoring of essay examinations, all the following are generally considered desirable 
   practices except to 
   A. reduce the mark for poor spelling or penmanship.  
   B. prepare a scoring key and standards in advance.  
   C. remove or cover pupils' names from the papers.  
   D. score one question on all papers before going to the next.

37. When is it generally desirable for the teacher to decide upon the specific format of items to 
   be developed for a test? 
   A. When the evaluation plan is being developed.  
   B. As the very first step.  
   C. After the total number of questions has been decided upon. 
   D. After study of the specific behaviors listed in the test plan.

38. One of the best ways for a teacher to begin a study designed to formulate goals for his 
   teaching is to 
   A. read the authors' prefaces of the textbooks he uses. 
   B. prepare an outline of the materials covered in his textbooks.  
   C. examine objectives formulated by other teachers. 
   D. discuss the problem with more experienced teachers.

39. The type of instructional outcome most difficult to evaluate objectively is 
   A. a concept. 
   B. an appreciation. 
   C. an attitude. 
   D. an understanding.

40. "Columbus discovered America in ________ ." 
   The best change to make in revising this item would be to rewrite it so as to read 
   A. "America was discovered by Columbus in ________ ." 
   B. "Columbus discovered ________ in ________ ." 
   C. "Columbus discovered America in the year of ________ ." 
   D. "_______ was discovered by Columbus in ________ ."

41. In which way are teacher-made tests superior to standardised tests? 
   A. They are more reliable for evaluating differences among very poor and very good students. 
   B. They provide more valid measures of the teacher's specific objectives. 
   C. They provide a better measure of the student's grasp of important facts and principles. 
   D. They are simpler to administer and score.
l2. This exercise
A. is faulty because the answers are not of parallel construction.
B. is faulty because the answers do not all complete the item stem.
C. is faulty because of ambiguous phraseology.
D. is faulty because the problem is not in the item stem.

l3. Measurement specialists would generally consider the practice of allowing a choice in the questions to be answered on an essay examination
A. desirable, because it gives each student a fairer chance.
B. desirable, because it permits a wider sampling of the topics covered.
C. undesirable, because it reduces the comparability of the test from student to student.
D. undesirable, because students waste too much time deciding which question to answer.

l4. A science teacher is preparing a test to be used to determine knowledge of specifics from a unit of study. He should use objective rather than essay questions because they
A. avoid ambiguity, the most common fault of test questions.
B. provide a wider sampling of material.
C. are not affected by the judgment of the tester.
D. are best suited to his purpose.

l5. One of the merits of arranging test items in an order of difficulty is that
A. it insures an accurate measure of consistency.
B. it encourages the pupil taking the test to continue.
C. item validity is to some extent dependent on difficulty.
D. this procedure contributes to the test's reliability.

For each of the following paired items, blacken one lettered space to indicate that the first item is
A. greater than the second
B. less than the second
C. definitely equal to the second
D. of uncertain size with reference to the second

l6. The level of ability represented by an I.Q. of 116 on the Stanford-Binet.

l7. The level of achievement in reading represented by a grade score of 8.5 on the California Reading Test.

l8. The justification of calling a test standardized that has been normed on 2,000 students.

l9. The desirability of using standardized achievement test results for grading purposes.

50. Extent to which correlation of parts is justified in a test designed to measure "general" intelligence.

After the number on the answer sheet which corresponds to that of each of the following items, blacken one lettered space to indicate the correct answer.

51. In determining the grade placement of pupils new to a school, the most useful data may be obtained by administering
A. achievement tests in reading, arithmetic and science.
B. achievement tests in reading and arithmetic.
C. achievement tests in reading and arithmetic plus an attitude inventory.
D. a survey achievement battery.
52. What is usually the last step in the production of a standardized achievement test?  
A. Final revision of test items and directions.  
B. Administration to a large and representative sample of pupils.  
C. Careful evaluation of test materials by experts.  
D. Statistical analysis of test items.

53. If you were asked to serve on a committee for the purpose of selecting a standardized achievement battery for your school, or school district, you would consider each of the following but give greatest weight to  
A. Unit cost per pupil tested.  
B. Availability of equivalent forms.  
C. Relevance to local instructional objectives.  
D. Ease of administration and scoring.

54. In a battery measuring various aptitudes the subtests should have  
A. Low correlations with each other and high reliability coefficients.  
B. High correlations with grade-point averages in college.  
C. Negative correlations with each other.  
D. Validity coefficients higher than their reliability coefficients.

55. In giving a standardized test a teacher allows too much time. This is most likely to adversely affect  
A. The reliability of the test.  
B. The validity of the test.  
C. Interpretation in terms of norms.  
D. The ranking of pupils.

56. Test techniques are generally preferred to observational techniques, when both are available for the testing purpose, because the former are  
A. More apt to yield measures.  
B. Perceived as a test by the student, thus more apt to be based on a motivated performance.  
C. Applicable to a wider variety of personal traits.  
D. More apt to yield reliable scores.

57. If, in administering a standardized test, one departs from the exact instructions, this will probably affect most seriously the  
A. Reliability of measurement.  
B. Objectivity of scoring.  
C. Applicability of norms.  
D. Comparability of individual scores.

58. Teachers should motivate students to make the best scores they possibly can on all of the following except  
A. Aptitude measures.  
B. Diagnostic measures.  
C. Personality measures.  
D. Readiness measures.

59. If a teacher wishes to obtain a critical review of a standardized test she plans to use with her classes, she should consult the  
A. Test Manual issued by the publisher.  
C. Review of Educational Research.  
D. Mental Measurements Yearbook.

60. In contrast to a test which is "well standardized" a poorly standardized test is one which  
A. Has norms that are based on fewer than 1,000 cases.  
B. Uses a norm sample that is not representative of the group for which the test is designed.  
C. Consists of test questions that have not been validated.  
D. Includes test questions that do not measure what they are intended to measure.

When you have finished the test, make sure that each mark on your answer sheet is solid, black, and glossy.  
Erase all superfluous marks on the answer sheet no matter how tiny.  
Then turn in your booklet and answer sheet to the proctor in charge.
APPENDIX C
1. What are some advantages and disadvantages of standardized tests?

   **Advantages**

   **Disadvantages**

2. Compare and contrast standardized tests and teacher-made tests from the point of view of uses in particular situations (for example, choice between one or the other type as a final examination in a course).

   **Standardized Tests**

   **Teacher-Made Tests**
3a. The Metropolitan Achievement Test Battery was given to all the children in a small city at the end of the 5th grade. The average grade level on the arithmetic computation test was 5.4. The superintendent of schools, examining the results, concluded that

I. his fifth graders fell below the national average in this test
II. his teachers were below average in teaching arithmetic
III. more time should be devoted to arithmetic in his schools

Which of these conclusions can safely be drawn from the facts reported?

A. None of them  
B. I only  
C. I and II only  
D. I and III only  
E. All of them

3b. On the Kuhlmann-Anderson Intelligence Test, John fell at the 98th percentile on the L (language) score and the 90th percentile on the Q (quantitative) score. Henry fell at the 55th and 45th percentiles on the same two scores. Who showed the greater unevenness in performance?

A. John  
B. Henry  
C. There was no difference.  
D. The data provide no basis for judging.

3c. When tested in the 11th grade with the Lorge-Thorndike Intelligence Test, a student got an IQ of 114. When tested in the 12th grade with the College Entrance Examination Board Scholastic Aptitude Test (SAT), he fell at the 48th percentile. How is the difference in these results best explained?

A. The student has not worked very hard in the senior year of high school.  
B. Anxiety probably depressed the student's score on the second test.  
C. The difference is merely that between IQ and percentile units.  
D. College Board norms are based on a very select sample.
4a. In administering a standardized achievement test, the examiner who paraphrases/elaborates on the test directions may

A. improve the validity of the test
B. improve the reliability of the test
C. invalidate the norms
D. both (A) and (B) above

4b. In administering a standardized achievement test, the examiner who allows a few extra minutes on the test may

A. improve the test validity for slower students
B. improve the test validity for all students
C. invalidate the norms

5a. If a student wanted to find critical reviews of the Iowa Tests of Basic Skills, he might best consult

A. Buros - The Mental Measurements Yearbooks
B. Measurement and Evaluation in Guidance
C. the manual of the test
D. Educational and Psychological Measurement

5b. Which of the following would be most useful for finding statistical information concerning the norms for a particular standardized test?

A. Journal of Consulting Psychology
B. the publisher's catalog
C. the manual of the test
D. a book on statistical methods in education

5c. If a student wanted to find out what validation studies had been done on the Kuder Preference Record in the last two or three years, he should go to

A. Buros - The Mental Measurements Yearbooks
B. Psychological Abstracts
C. Encyclopedia of Educational Research
D. Review of Educational Research
5d. If a student wanted to find out the most recently published achievement tests in arithmetic, he might best consult

A. publishers' catalogs  
B. Buros - The Mental Measurements Yearbooks  
C. Journal of Educational Measurement  
D. Tests in Print

6. It would be most accurate to say that intelligence tests measure

A. a sample of the behavior of an individual  
B. the innate capacity of an individual  
C. the maturity level of an individual  
D. the probable future success of an individual

7. A major difference between the Stanford-Binet and the Wechsler series of intelligence scales is that

A. the Stanford-Binet yields a result expressed as an IQ  
B. the Wechsler tests are suitable for group administration  
C. the Stanford-Binet is based quite directly on school learnings  
D. the Wechsler tests are made up of separate subscales that yield separate scores
For the following statements (Questions 6&7a--6&7g) mark:

A. if the statement applies more to group intelligence tests
B. if the statement applies more to individual intelligence tests
C. if the statement applies about equally to both types

6&7a. Provide opportunity for clinical and diagnostic observations of individual behavior

6&7b. Require that the subject be motivated to do his best

6&7c. Handicap those with poor reading skills

6&7d. Provide little opportunity to appraise originality of response or inventiveness

6&7e. Require judgment and some skill to score

6&7f. Have substantial reliability

6&7g. Suitable for use with children of preschool age

8a. In which of the following would an adjustment inventory be most likely to be effective?

A. In screening applicants for employment as executives in a business concern
B. In picking out elementary-school children for a special educational program
C. For studying adjustment problems in a prison population
D. As a first step in evaluating the problems of students who had come to a clinic for personal counseling
8b. The main difficulty that a classroom teacher would find in using the Minnesota Multiphasic Personality Inventory (MMPI) would be that

A. only two forms of the test exist
B. raw scores cannot be transformed into scale scores
C. the questions are weighted
D. clinical training is necessary to interpret score profiles

8c. What was the criterion by which the scoring keys for the Strong Vocational Interest Blank (SVIB) were determined? Weights were assigned to items which

A. appeared logically to belong together and form a common pattern
B. distinguished successful from unsuccessful workers in an occupation
C. distinguished successful men in an occupation from men in general
D. correlated with a group of other items

8d. If an interest inventory were included in a high-school guidance program, the most satisfactory way to use the results of the inventory would be for the counselor to

A. report the scores to the student and let him interpret them
B. study the scores and report to the student the fields for which he seems best qualified
C. use the test scores as a basis for an interview to explore the student's interests and plans
D. prepare a written report to be sent to the student's parents

9a. The basic assumption of each of the projective methods is that the responses that an individual makes to the stimulus materials depend primarily upon the

A. nature of the stimulus presented
B. individual's previous experience with the stimuli
C. individual's inner personality structure
D. individual's present mood and feeling tone
9b. Up to the present time, the Rorschach Inkblot Test has been used primarily for

A. predicting success in college
B. predicting success in various vocational fields
C. determining whether a person is using his abilities to the fullest extent
D. clinical diagnosis

9c. A review of studies that have undertaken to investigate the validity of the Rorschach Inkblot Test leads to the conclusion that

A. when the study has been adequately designed, the Rorschach has been shown to be valid in almost every case
B. the Rorschach test has validity primarily for predictive purposes
C. evidence for validity has been spotty, with many negative results
D. the evidence on validity has been very largely negative, and there is little evidence to support the claims for this test

9d. The greatest limitation on the use of projective techniques in education is that they

A. require highly skilled administrators
B. are not reliable
C. are scored subjectively
D. require too much time to administer

10a. The type of examination that would be of greatest value for instructional purposes in colleges would be

A. a comprehensive examination between the sophomore and junior years
B. a final examination at the end of a course
C. an examination at the end of each unit of subject matter
D. a short quiz given once a week
10b. Which of the following can provide the most effective direction and guidance of a pupil's learning?

A. A monthly grade in each subject
B. A comprehensive year-end examination
C. Prompt analysis of his errors, and a report of them to him
D. Opportunity for parent-teacher conferences

10c. In the use of tests and other evaluative techniques in the classroom, the highest priority should be given to

A. assigning course grades
B. improving instructional decisions
C. maintaining adequate school records
D. reporting pupil progress to parents

END OF PART I—STANDARDIZED TESTS.

CHECK YOUR WORK AGAINST THE ANSWER KEY.
11. What are some advantages and disadvantages of teacher-made tests?

**Advantages**

**Disadvantages**

12a. Probably the best way to plan a classroom test is to make a

A. two-way grid  
B. statistical analysis  
C. taxonomy  
D. list of instructional objectives  
E. list of what students consider most important

12b. The most important requirement of a test item is that it

A. measure a specific behavior  
B. discriminate between good and poor students  
C. challenge the student to think  
D. measure achievement of a teaching objective  
E. be unambiguous
13a. Which one of the following is an example of a behavior term?
   A. Fears
   B. Identifies
   C. Realizes
   D. Thinks

Following is a list of statements that a teacher compiled to clarify what he meant by understanding principles. If a statement is properly stated in behavioral terms, mark B. If it is not properly stated in behavioral terms, mark N.

   B - Behavioral
   N - Nonbehavioral

   ____ 13b. Sees the value of the principle
   ____ 13c. Makes a prediction using the principle
   ____ 13d. Describes situations in which the principle is applicable
   ____ 13e. Realizes the essential features of the principle
   ____ 13f. Is familiar with the uses of the principle
   ____ 13g. Identifies misapplications of the principle
   ____ 13h. Explains the principle in his own words
   ____ 13i. States tenable hypotheses based on the principle
   ____ 13j. Appreciates the complexity of the principle
   ____ 13k. Develops a complete understanding of the principle

14a. When should a two-way grid be prepared?

   A. Just before the test is given
   B. Before the test items are written
   C. As the test items are written
   D. After the test is given
   E. As the test is scored
14b. Which of the following questions is most important in evaluating a test item?

A. Is the keyed response factually correct?
B. Are the distractors plausible yet incorrect?
C. Is the item difficult enough?
D. Is there a specific determiner in the item?
E. Should students be expected to answer it correctly?

15a. The most serious limitation of the multiple choice type of item is that it

A. cannot appraise originality
B. requires a high level of reading skill
C. is limited to the appraisal of recall of knowledge
D. encourages guessing

15b. Most of the research evidence on the relationship between performance on a free-answer test and an objective test indicates that the relationship is

A. negligible; i.e., just about zero
B. positive but low
C. positive and high
D. negative but low
E. negative and high

15c. What is an advantage of objective tests in comparison with essay tests?

A. Greater validity
B. Less expensive
C. Easier to prepare
D. Easier to score
E. Easier to administer
For items 15d--15f, select the one letter from the key to indicate the type or types of test items being referred to:

Key:  
A. True-false items  
B. Multiple-choice items  
C. Matching items  
D. Essay items  
E. More than one of the above types

____ 15d. The answers completing the items should be of parallel construction.

____ 15e. Such items are more useful than other types in measuring students' knowledge of definitions of a number of technical terms.

____ 15f. Such items can often be made more effective by allowing students, under certain conditions, to change a word or phrase in the item.

16a. To properly administer a standardized test, it is necessary to

A. have all directions listed in the test itself  
B. give recognition to any factors peculiar to the group tested  
C. give pupils all necessary directions before starting the test  
D. study thoroughly and follow carefully the printed instructions  
E. avoid detailed and highly verbal directions

16b. The primary function of a proctor is to

A. interpret ambiguous test items to individuals who raise questions  
B. maintain testing conditions as defined in the test manual  
C. supplement the test directions where this seems necessary  
D. make sure that all individuals finish the test within the time limits allowed  
E. be prepared to take over for the examiner, if necessary
16c. In administering a standardized test, it is best to

A. discourage entirely the asking of any questions
B. answer questions about test items only when necessary
C. answer questions only about procedure once the test has begun
D. permit no questions after the test is started
E. require pupils to come to the teacher's desk to ask questions

17a. Which of the following is the most appropriate statement about good essay questions?

A. They should be stated briefly and clearly.
B. Extent of treatment required should be given.
C. They should reflect instructional objectives.
D. All of the above.
E. Both (A) and (C) above.

17b. Which of the following is not considered a sound principle in constructing true-false items?

A. Avoid statements which might be misinterpreted.
B. Confine each item to a single idea.
C. Avoid statements that are long and complex.
D. Use more true than false statements.
E. Avoid negative statements as much as possible.
In Questions 17c--17g, a test item for a final examination for 7th-grade general science, and the purpose of the questions, are given. After each item, three suggested changes are given for improving the question. Read each suggested change and then select the one, if any, that would make the item or test technically better. If none of the suggested changes would improve the item or test, mark D. (Note that the correct answer to each test item is marked by an asterisk, *.)

17c. *T  F All bacteria are pathogenic. (Purpose: to measure the meaning of pathogenic.)

A. Many bacteria are pathogenic.
B. Rewrite as a completion item: Bacteria are ________________.
C. Rewrite as a multiple-choice item: An organism that causes disease is said to be (1) *pathogenic; (2) saprophytic; (3) antigenic; (4) pandemic.
D. None of the suggested changes improves the item.

17d. Column I

____ Digestive enzyme
____ Sense organ
____ Noted scientist
____ Part of the tooth

Column II

1. Pasteur
2. Pulp
3. Lipase
4. Ear

*(Answers 3, 4, 1, 2) (Purpose: to test knowledge of specific facts.)

A. Rewrite as four separate short-answer items.
B. Add more items in Column II and arrange Column II in alphabetical order.
C. Make both Column I and Column II longer.
D. None of the suggested changes would improve the item.

17e. Which of the following is not a digestive enzyme? (1) ptyalin; (2) *thyroxin; (3) erepsin; (4) maltose. (Purpose: to test knowledge of what thyroxin is.)

A. Rewrite item to read: Thyroxin is an example of (1) a digestive enzyme; (2) an end-product of digestion; (3) a hormone; (4) a waste material of metabolism.
B. Rewrite as a completion item: Thyroxin is ________________.
C. Rewrite as a true-false item: Thyroxin is not a digestive enzyme.
D. None of the suggested changes improves the item.
17f. *T F People should not smoke cigarettes. (Purpose: to test attitudes toward smoking.)

A. Reword statement to read: Most medical authorities state that people should not smoke.
B. Rewrite as a completion item: People should not smoke ______
C. Rewrite as a multiple-choice item: People should not smoke cigarettes because smoking (1) causes cancer of the lungs; (2) is expensive; (3) is habit-forming; (4) causes allergies.
D. None of the suggested changes improves the item.

17g. The scientific name for the eardrum is (1) *tympanum; (2) epidermis; (3) duodenum; (4) stirpes. (Purpose: to test knowledge of scientific terminology.)

A. Rewrite as a true-false item: The scientific name for the eardrum is septum.
B. Rewrite as a completion item: The name for the eardrum is ______.
C. Rephrase item as a short answer question: What is the technical term for the eardrum?
D. None of the suggested changes improves the item.

18a. If a group of students attempted to guess the correct responses on all the items of a true-false test, an approximation of the average percent of correct responses would be

A. 10
B. 25
C. 50
D. 75
E. impossible to estimate

18b. In a multiple-choice examination made up of four-choice items, if one wanted to correct the results for guessing he would be most likely to score the examination

A. rights minus 1/4 wrongs
B. rights minus 1/3 wrongs
C. rights minus 1/2 wrongs
D. rights minus wrongs
18c. Why would it be undesirable to correct for guessing on the typical classroom test?

A. Correction for guessing would overpenalize the bold guesser.
B. Pupils' guesses are frequently blind guesses.
C. Pupils' guesses are usually informed guesses.
D. Pupils seldom guess on classroom tests.

19. It has sometimes been proposed that in scoring an essay examination in a subject such as science,

I. a sample answer be prepared for each question before any grading is done
II. all papers be scored on a single question before going on to the next
III. the grades take account of mechanics of writing, as well as the ideas included

Which of these are desirable procedures?

A. Only I
B. Only I and II
C. Only II and III
D. Only I and III
E. I, II, and III

20. What use should be made of standardized test results in discussions with the parents of an elementary-school pupil?

A. As a rule, the parent should be told the child's exact scores on any standardized tests.
B. Test results should be reported only in special cases, but should then be reported exactly as they appear in school records.
C. An interpretation of the test results should be given, but scores should usually not be reported.
D. No mention should be made of standardized test results.
21a. A good marking and reporting system will be based on

   A. a single letter grade that represents achievement, effort, and attitude
   B. objective test scores only
   C. the normal curve
   D. the objectives of the school

21b. Which of the following is an acceptable reason for grading on the curve as opposed to other methods?

   A. The grading process is objective.
   B. The process spreads the scores evenly.
   C. The process is based on statistical principles.
   D. The process is entirely fair.
   E. None of the above.

22a. One of the most serious faults of the essay test is that it

   A. often is graded on irrelevant factors such as quality of handwriting
   B. uses an absolute rather than a relative scale
   C. involves too much writing on the part of the pupils
   D. cannot be adapted to standardized testing procedures
   E. requires too much clerical help

22b. One advantage of essay questions over objective items is that they

   A. have higher validity
   B. minimize guessing
   C. provide for more adequate sampling
   D. provide for more consistent scoring

22c. Essay questions are more useful than objective items for measuring the

   A. application of principles
   B. integration of factual information
   C. interpretation of data
   D. recall of learned material
23a. A table of specifications, or blueprint, is used in test construction primarily to ensure more adequate
A. arrangement of items
B. control of item difficulty
C. sampling of content
D. scoring of results

23b. A table of specifications, or blueprint, would be useful for all of the following EXCEPT
A. the construction of a classroom test
B. the construction of a rating scale
C. the development of a general evaluation plan
D. the selection of a standardized test

END OF PART II--CONSTRUCTION AND EVALUATION OF CLASSROOM TESTS.
CHECK YOUR WORK AGAINST THE ANSWER KEY.
24a. If a teacher wished to do diagnostic testing to determine whether many of the pupils in her class were having special difficulty with the *ie* and *ei* spelling combination, it would be best for her to

A. give a standardized spelling test
B. prepare her own test, with many words involving these combinations
C. test each child orally, to see how he spelled the words
D. have the pupils write compositions, and make a count of the errors in these

24b. A survey achievement test would provide a more valid measure of a pupil's level of achievement than a diagnostic test because

A. it contains items that are more difficult
B. it is easier to administer and score
C. it typically has more subtests
D. it uses objective test items only

24c. The main purpose of a diagnostic test is to determine

A. the effectiveness of the school curriculum
B. the pupils' level of learning in relation to national norms
C. the type of remedial work needed
D. which pupils should be passed and which should be failed

24d. A diagnostic test is one test that is used primarily to

A. predict future progress along a certain line
B. determine strengths and weaknesses in achievement
C. estimate the likelihood of success in higher levels of education
D. lay a foundation for psychological counseling and guidance
E. form homogeneous groups within the classroom
24e. What is the only kind of standardized test for which it is always legitimate to discuss with a pupil his answers to specific items?

A. Diagnostic achievement test
B. General scholastic aptitude test
C. Individual intelligence test
D. Test in a survey battery
E. Group intelligence test

25a. Juanita is exactly 10 years old. Her mental age is 11 years 6 months. What is her ratio IQ?

A. 87
B. 106
C. 115
D. 126

25b. A nine-year-old child has an IQ of 120. What is her mental age to the closest year?

A. 8
B. 9
C. 11
D. 12
E. 15

26. The range of intelligence generally classified by most authorities as "normal" or "average" includes persons having IQs which are

A. about 100
B. in the 95–105 range
C. in the 90–110 range
D. in the 80–120 range
E. in the 70–130 range
A teacher noted that two of his pupils had the following percentile ranks on the verbal and numerical ability subtests of an aptitude battery.

<table>
<thead>
<tr>
<th></th>
<th>Verbal Ability</th>
<th>Numerical Ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann</td>
<td>98</td>
<td>84</td>
</tr>
<tr>
<td>Harry</td>
<td>50</td>
<td>36</td>
</tr>
</tbody>
</table>

Assume a normal distribution of scores and answer questions 27a and 27b.

27a. Who showed the greater difference in performance on the two tests?
   A. Ann
   B. Harry
   C. The differences in performance are equal.
   D. More data are needed to determine this.

27b. On which test did the performance of Ann and Harry differ most?
   A. Verbal ability
   B. Numerical ability
   C. The differences in performance are equal.
   D. More data are needed to determine this.

28. What is a limitation of a mathematics test which requires the reading of material in order to solve a problem?

29a. The fairest percent correct for a passing grade is
   A. 75
   B. 70
   C. 65
   D. 60
   E. impossible to say
29b. What is a limitation of the "percentage" system of marking, i.e., assigning grades on a specified percent of correct answers?

A. The system is arbitrary.
B. The system generally does not reflect clearly-defined performance standards.
C. Both A and B.
D. There is no limitation in such a procedure.

30a. If percentile norms are to provide a meaningful picture of an individual's performance, they must be

A. expressed in equal units of score
B. translated into quotient values
C. revised every year or two
D. based on a group of which he may be considered a member

30b. Why is it not sensible to expect the typical school system to bring all fourth graders up to the fourth-grade norm on a standardized achievement test in such a subject as reading? Because the norm

A. is an average, rather than a minimum standard
B. is designed for average and above-average children
C. takes no account of limited social and cultural background
D. has been moved up in recent years

31. The same examination was given in Class I and Class II.

<table>
<thead>
<tr>
<th>Class</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class I</td>
<td>49.5</td>
<td>10.5</td>
</tr>
<tr>
<td>Class II</td>
<td>55.5</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Barbara earned a score of 60 on the examination. Her score would rank relatively higher with respect to which class?

A. Class I
B. Class II
C. Equally high with Class I and Class II
D. The answer cannot be determined from the information given.
Questions 32a--32g are based on statements that could have appeared in the validity sections of various test manuals. Read each statement and describe what kind of validity evidence is represented by the statement. For your answer, mark

A. for content validity  
B. for predictive validity  
C. for construct validity

32a. The items on the Machinist Proficiency Test are based on an analysis of the job of machinists in 100 plants that employ five or more machinists.

32b. Scores on the Alpha Group Intelligence Test given to high school seniors were correlated with grades at the end of freshman year in college.

32c. The Gates Reading Test given in May of first grade was correlated with the Kuhlmann-Anderson Intelligence Test given at the end of the same school year.

32d. Scores on the Social Adjustment Inventory were correlated with ratings of leadership given by teachers at the time that the test was given.

32e. An inventory of study skills and habits was correlated with grade-point average obtained one year later.

32f. Classroom teachers rated each item on the Arithmetic Reasoning Test (Junior High School Level) on a five-point scale for significance and importance in the junior high school mathematics curriculum.

32g. The correlation between scores on the Sales Selection Inventory and sales records was +.61.

32h. From the reliability coefficient of a test, one can judge

A. how many points a pupil is likely to change if an equivalent test is given  
B. how consistently a pupil will maintain his position in the group if an equivalent test is given  
C. whether the test is measuring what it is supposed to measure  
D. whether the test is related to other significant factors in the individual
32i. What procedure gives the most rigorous and exacting definition of reliability?

A. Retesting with the same test a month later
B. Subdividing test items into two halves, and correlating half-test scores
C. Administering an equivalent form of the test a month later
D. Procedures A, B, and C are logically equivalent.

32j. The justification for estimating reliability by the split-half procedure is that this procedure

A. is convenient
B. gives higher reliability coefficients
C. involves the smallest number of assumptions
D. can be used appropriately for speeded tests

32k. The standard error of measurement is best described as the standard deviation of a distribution of

A. scores for a homogeneous group
B. scores for a single form of a test
C. differences between scores on two testings
D. repeated measurements of a single individual

32l. Which of the following statements could not possibly be true for an aptitude or achievement test?

A. Though it has little face validity, it shows substantial statistical validity.
B. Though it is judged to have high content validity, it has very low reliability.
C. Though it has zero reliability, it has substantial statistical validity.
D. Though it has zero statistical validity, its reliability is quite high.
32m. **Item difficulty** refers to the percentage of a given group who
   A. answer the item correctly
   B. answer the item incorrectly
   C. attempt to answer the item
   D. leave the item blank

32n. The **index of discrimination** of an item refers to
   A. the proportion of high scorers passing the item
   B. the proportion of high scorers selecting each alternative
   C. the proportion of low scorers passing the item
   D. the proportion of high scorers passing the item minus the proportion of low scorers passing the item

32o. A distractor is judged good if it attracts
   A. a majority of all pupils answering the item
   B. an approximately equal number of high achieving and low achieving pupils
   C. more high achieving pupils
   D. more low achieving pupils

33. From previous tests, a teacher has four items that would be suitable for Test A. She plans to use only one of these items. On the basis of the item analysis below, which item should she choose? (Correct answer shown by asterisk for each teacher-made item.)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B*</td>
<td>C</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>High Group</td>
<td>0</td>
<td>96%</td>
</tr>
<tr>
<td>Low Group</td>
<td>8%</td>
<td>80%</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C*</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>High Group</td>
<td>12%</td>
<td>36%</td>
</tr>
<tr>
<td>Low Group</td>
<td>20%</td>
<td>60%</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C*</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>High Group</td>
<td>60%</td>
<td>16%</td>
</tr>
<tr>
<td>Low Group</td>
<td>16%</td>
<td>24%</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A*</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>High Group</td>
<td>84%</td>
<td>0</td>
</tr>
<tr>
<td>Low Group</td>
<td>68%</td>
<td>0</td>
</tr>
</tbody>
</table>
34. What is a limitation of ability grouping based on only one measure of ability?
A. An ability test needs to be supplemented by achievement tests.
B. An ability test needs to be supplemented by the teacher's observations.
C. An ability test needs to be supplemented by other ability tests.
D. All of the above.

35a. It would be most accurate to say that intelligence tests measure
A. a sample of the behavior of the individual
B. the innate capacity of an individual
C. the maturity level of an individual
D. the probable future success of an individual

35b. As contrasted with the measurement of height or weight, the measurement of intelligence is
A. absolute
B. impossible
C. indirect
D. precise
E. qualitative

36a. If a pupil ranks ninth (first is best) in a class of 30, his percentile rank is
A. 27
B. 30
C. 33
D. 70
E. 75

36b. The major purpose of a frequency distribution is to
A. determine the correlation between two sets of data
B. make predictions about a set of data
C. describe a set of data
D. determine the reliability of a set of data
36c. Between which two percentile ranks is there probably the most difference in ability represented?

A. 1 and 2  
B. 25 and 26  
C. 49 and 51  
D. 75 and 76  
E. 90 and 91

END OF PART III--USES OF MEASUREMENT AND EVALUATION.

CHECK YOUR WORK AGAINST THE ANSWER KEY.
37a. The following scores were obtained by 15 individuals on a test of spatial relations:

48, 20, 36, 38, 19, 42, 46, 33, 41, 21, 37, 50, 18, 28, 44

Rank these scores from 1 to 15, where 1 = lowest score.

37b. The following scores were obtained by 10 individuals on a personnel selection test:

24, 26, 19, 16, 17, 9, 12, 5, 12, 11

Rank these scores; 1 = highest score. Use the average-rank method for tied scores.

38a. Here are scores made by 100 students on a test:

| 69 | 59 | 55 | 53 | 52 | 50 | 49 | 45 | 43 | 40 |
| 64 | 58 | 55 | 53 | 52 | 50 | 48 | 45 | 43 | 40 |
| 63 | 58 | 55 | 53 | 52 | 50 | 47 | 45 | 42 | 37 |
| 63 | 57 | 54 | 53 | 52 | 49 | 47 | 45 | 42 | 37 |
| 63 | 57 | 54 | 53 | 52 | 49 | 47 | 45 | 42 | 36 |
| 62 | 57 | 54 | 53 | 52 | 49 | 47 | 45 | 42 | 35 |
| 61 | 57 | 54 | 52 | 52 | 49 | 47 | 45 | 42 | 33 |
| 61 | 56 | 54 | 52 | 51 | 49 | 46 | 44 | 41 | 32 |
| 61 | 56 | 53 | 52 | 51 | 49 | 46 | 43 | 41 | 31 |
| 60 | 55 | 53 | 52 | 51 | 49 | 46 | 43 | 40 | 31 |

In constructing a frequency distribution, what would be the size of the class interval, \(i\)?

A. 2  
B. 3  
C. 4  
D. 7
38b. You have IQs for 350 children and are preparing to make a frequency distribution. The IQs range from 62 to 134. Which would be the most satisfactory way to group the scores?

A. 62-63, 64-65, 66-67, etc.
B. 61-63, 64-66, 67-69, etc.
C. 60-64, 65-69, 70-74, etc.
D. 60-69, 70-79, 80-89, etc.

39a. The standard error of measurement is best described as the standard deviation of a distribution of

A. repeated measurements of a single individual
B. scores for a homogeneous group
C. scores for a single form of a test
D. differences between scores on two testings

39b. An individual's score on an achievement test is 75. The standard error of measurement for the test is reported to be 5 points. What are the chances that the individual's true score is between 70 and 80?

A. About 9 chances in 10
B. About 2 chances in 3
C. About 1 chance in 3
D. About 1 chance in 6

40. Histograms and frequency polygons are

A. inferential statistics
B. measures of variability
C. methods of graphical representation of data
D. geometric representations of the correlation coefficient
In the blank before each numbered item, write the letter that refers to the measure of central tendency most appropriate to the case in point.

A. Mean
B. Median
C. Mode

41a. The point on the scale at which the greatest number of cases fall

41b. The "average" determined by dividing the sum of scores by the number of cases

41c. The "average" which is equivalent to the 50th percentile

41d. The most commonly used "average"

42a. Here is a set of 8 measurements:

25 27 20 28 24 29 22 20

Find the mean, the median, and the mode.

42b. In the group of scores 1, 3, 3, 3, and 5 it can be said of the mean, the median, and the mode that

A. the mean is larger than either the median or the mode
B. the mode is larger than either the median or the mean
C. the median is larger than either the mean or the mode
D. all are different
E. all are the same
43a. The measure of central tendency most affected by extreme scores in a single distribution is the

A. mean
B. median
C. mid-score
D. mode

43b. In a country where comparatively few people earn extremely high incomes, the mean income will

A. be higher than the median and the mode
B. be lower than the median and the mode
C. be the same as the mode
D. be the same as the median
E. fall between the median and the mode

44a. A measure of variability provides information about

A. level of performance
B. shape of the distribution
C. both A and B above
D. neither A nor B above

44b. In a high school test, a teacher gave two sections of a class the same algebra test. The results were as follows:

Section I: Mean 48, Standard deviation 6.3
Section II: Mean 48, Standard deviation 3.2

Which of the following conclusions is correct?

A. Section I is more homogeneous than Section II.
B. Section II is more homogeneous than Section I.
C. Both sections are equally homogeneous.
D. Section II has brighter students than Section I.
In each of the items below, choose from the right-hand column the letter corresponding to the term which best matches the item in the left-hand column.

<table>
<thead>
<tr>
<th>Situation</th>
<th>Measures of Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>44c. When a quick approximation of variability is sought</td>
<td>A. Semi-interquartile range</td>
</tr>
<tr>
<td>45. When the degree of concentration around the median is sought</td>
<td>B. Range</td>
</tr>
<tr>
<td>46. When it is desirable to compute z-scores</td>
<td>C. Standard deviation</td>
</tr>
<tr>
<td></td>
<td>D. Average deviation</td>
</tr>
</tbody>
</table>

47. Find the semi-interquartile range of the following frequency distribution:

<table>
<thead>
<tr>
<th>Scores</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>27-29</td>
<td>1</td>
</tr>
<tr>
<td>24-26</td>
<td>2</td>
</tr>
<tr>
<td>21-23</td>
<td>4</td>
</tr>
<tr>
<td>18-20</td>
<td>5</td>
</tr>
<tr>
<td>15-17</td>
<td>3</td>
</tr>
<tr>
<td>12-14</td>
<td>2</td>
</tr>
<tr>
<td>9-11</td>
<td>2</td>
</tr>
<tr>
<td>6-8</td>
<td>1</td>
</tr>
</tbody>
</table>

48a. If a student obtains a standard z score of +2, he falls at approximately what percentile rank?
48b. Complete the following table.

<table>
<thead>
<tr>
<th>Standard z score</th>
<th>Approximate Percentile Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>+2</td>
<td></td>
</tr>
<tr>
<td>+3</td>
<td></td>
</tr>
</tbody>
</table>

49a. In a normal distribution, approximately 68 percent of the cases fall between

A. ±3 standard deviations from the mean
B. ±2 standard deviations from the mean
C. ±1 standard deviation from the mean
D. ±1/2 standard deviation from the mean
E. the median and the mean

49b. Complete the following table.

<table>
<thead>
<tr>
<th>Area Under Normal Curve</th>
<th>Percentage of Cases In Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>±1 standard deviation from the mean</td>
<td></td>
</tr>
<tr>
<td>±2 standard deviations from the mean</td>
<td></td>
</tr>
<tr>
<td>±3 standard deviations from the mean</td>
<td></td>
</tr>
</tbody>
</table>

50. The normal curve is best viewed as

A. an exact description of many kinds of data
B. a law of nature
C. a statistical model
D. all of the above
51. Standard scores obtained from different distributions may be compared

A. if the norm groups are comparable
B. if the shapes of the distributions are similar
C. if both A and B are true
D. irrespective of norm group and shape of distributions

52a. If a student obtains a score of 75 in a group where the mean is 84 and the standard deviation is 6, he falls

A. two standard deviations above the mean
B. two standard deviations below the mean
C. one-and-one-half standard deviations above the mean
D. one-and-one-half standard deviations below the mean

52b. A student scores 45 on a vocabulary test. The mean for the class is 47 and the standard deviation is 8. His z-score is

A. 0.25
B. 0.20
C. -0.20
D. -0.25

53. What is the mean and standard deviation of the following common standard score scales?

<table>
<thead>
<tr>
<th>Standard Score Scale</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>z</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stanine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>deviation IQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CEEB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
54. Given the following information, on which test has Johnny done the best? (Assume that all four tests have approximately the same "shape" of distribution.)

<table>
<thead>
<tr>
<th>Test</th>
<th>Johnny</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>4.5</td>
<td>6.0</td>
<td>1.5</td>
</tr>
<tr>
<td>Math</td>
<td>10.0</td>
<td>9.8</td>
<td>0.1</td>
</tr>
<tr>
<td>Science</td>
<td>7.1</td>
<td>6.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Phys. Ed.</td>
<td>16.0</td>
<td>15.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>

A. English
B. Mathematics
C. Science
D. Physical Education
E. Impossible to tell

55a. A student receives a z-score of -1.35 on a spelling test. What is his equivalent T-score?

55b. A university testing center had an established policy of converting all raw test scores for students into standard scores with a mean of 500 and a standard deviation of 100 (transformation A). The computation center of the university recently requested the research center to change the standard-score system to one with a mean of 5 and a standard deviation of 2 (transformation B) so that the transformed scores could be placed in a single column on an IBM card, permitting more efficient IBM machine computation.

Convert the following transformation A scores to transformation B scores:

550 400 600 500
56. The mode, mean, and median are the same

A. always
B. never
C. when the distribution of scores is symmetrical
D. when the distribution of scores is not symmetrical

In each of the items below, choose from the right-hand column the letter corresponding to the term which best matches the item on the left.

57a. A. positively-skewed distribution
     B. negatively-skewed distribution
     C. bimodal distribution
     D. dichotomous distribution

57b. A.

57c. A.

58. Any normal distribution can be completely described in terms of its

A. mean
B. standard deviation
C. skewness
D. both A and B above
E. A, B, and C above

59a. Define correlation coefficient.
59b. Define positive correlation; negative correlation; no relationship; perfect relationship.

Below are listed five correlation coefficients.

A. .85
B. .50
C. .00
D. -.63
E. -.92

Questions 60a and 60b refer to these choices.

_____ 60a. Which of the five would permit the most accurate prediction?
_____ 60b. Which of the five would indicate that two tests were measuring two unrelated skills?

60c. If \( r = -1.00 \),

A. \( X \) is of no use in predicting \( Y \)
B. values of \( Y \) can be errorlessly predicted from values of \( X \)
C. the mean of \( X \) is lower than the mean of \( Y \)
D. the standard deviation of \( X \) is smaller than the standard deviation of \( Y \)
61. A high positive correlation between excess weight and heart attacks shows that

A. there is a common cause for excess weight and heart attacks
B. loss of weight will reduce the possibility of heart attacks
C. excess weight does not cause heart attacks
D. none of the above

62. A correlation coefficient of -0.90 between two tests means that

A. 90% of the variance of test 1 is accounted for by test 2
B. 81% of the variance of test 2 is accounted for by test 1
C. 90% of the high scores on test 1 go with low scores on test 2
D. 81% of the low scores on test 1 go with high scores on test 2

63. What range of correlation coefficients would you classify as "high," "low," and "moderate"?

| High: From _____ to _____ |
| Low: From _____ to _____ |
| Moderate: From _____ to _____ |

In each of the items below, choose from the right-hand column the letter corresponding to the term which best matches the item on the left.

A. positive correlation \((r = +)\)
B. negative correlation \((r = -)\)
C. perfect relationship \((r = 1.00)\)
D. no relationship \((r = 0.00)\)
What size of correlation would you expect for the following items?

____ 65a. Stanford-Binet and Wechsler IQ tests
   A. high positive correlation
   B. low positive correlation
   C. approximately zero correlation
   D. low negative correlation
   E. high negative correlation

____ 65b. Number of storks present and number of babies born
   A. high positive correlation
   B. low positive correlation
   C. approximately zero correlation
   D. low negative correlation
   E. high negative correlation

____ 65c. IQ score and job success rating

66. If a student has a score of 50 on Test I and a score of 100 on Test II, it can be said that

   A. he did twice as well on Test II as on Test I
   B. Test II was the easier of the two tests for him
   C. he achieved a perfect score on Test II
   D. both A and C above
   E. there is no basis for comparing the scores as given here

67. Which of the following kinds of scores indicates the distance from the mean least accurately?

   A. Grade-equivalent score
   B. Deviation IQ
   C. Percentile rank
   D. Standard score
   E. Stanine score
For each of the following instances, state the highest level of measurement scale involved. For questions 68a--68f, write I. for interval scale
   N. for nominal scale
   O. for ordinal scale
   R. for ratio scale

___ 68a. Numbers of men and women in a tests and measurements class

___ 68b. Number of pounds that a person can lift

___ 68c. Temperature on a Celsius scale

___ 68d. Numbers assigned consecutively to students as they complete an exam

___ 68e. Numbers assigned to four kinds of cola drinks

___ 68f. Ranking of five students from best to worst in terms of potential for graduate study

---

69a. Michele's history test raw score of 35 corresponds to a percentile rank of 60. This means that

A. Michele got 35% of the history test items correct
B. Michele's score is higher than 60% of the scores in the norm group
C. Michele's score is lower than 60% of the scores in the norm group
D. Both A and B above are correct

69b. Between which two percentile ranks is there probably the most difference in ability represented?

A. 1 and 2
B. 25 and 26
C. 49 and 51
D. 75 and 76
E. 90 and 91
Interpretations of achievement from norms is affected by

A. ability level
B. cultural background
C. curricular factors
D. all of the above
E. none of the above

END OF PART IV--STATISTICAL CONCEPTS.

CHECK YOUR WORK AGAINST THE ANSWER KEY.
APPENDIX D
ANSWER KEY
DIAGNOSIS OF MEASUREMENT
COMPETENCY
PART I--STANDARDIZED TESTS

Directions: Check your test answers against this Answer Key. Circle the item numbers for every incorrect answer. Then refer to the Prescription Sheet to identify the corresponding learning objectives and the prescriptive page references.

1. **Advantages**
   Broad coverage; rigidly controlled procedure of administering and scoring; availability of norms for evaluating scores; high quality of test items.

   **Disadvantages**
   Inflexibility for evaluating learning outcomes unique to particular school, class, or content area.

   (Similar or related answers acceptable)

2. **Standardized Tests**
   (1) Situations in which comparisons need to be made.
   (2) Situations in which there are large numbers of people about whom decisions need to be made, but for whom the decision maker has no common or comparable data.

   **Teacher-Made Tests**
   (1) Mastery of limited unit of instruction.
   (2) Achievement of distinctive local objectives.
   (3) Assigning of marks.

   (Similar or related answers acceptable)

3a. B
3b. A
3c. D

4a. C
4b. C
5a. A
5b. C
5c. B
5d. A

6. A

7. D

6&7a. B
6&7b. C
6&7c. A
6&7d. A
6&7e. B
6&7f. C
6&7g. B

8a. D
8b. D
8c. C
8d. C

9a. C
9b. D
9c. C
9d. A

10a. D
10b. C
10c. B

NOW REFER TO PRESCRIPTION SHEET.
FOR EACH INCORRECT ITEM, IDENTIFY CORRESPONDING LEARNING OBJECTIVE.
THEN REFER TO PRESCRIPTIVE PAGE REFERENCES IN YOUR TEXTBOOK.
ANSWER KEY

DIAGNOSIS OF MEASUREMENT COMPETENCY

PART II--CONSTRUCTION AND EVALUATION OF CLASSROOM TESTS

Directions: Check your test answers against this Answer Key. Circle the item numbers for every incorrect answer. Then refer to the Prescription Sheet to identify the corresponding learning objectives and the prescriptive page references.

11. Advantages
   Geared to outcomes and content of local curriculum; flexible to adapt measurement to new materials and changes in procedure.

Disadvantages
   Quality of test items often low or unknown; comparison to norm group not usually possible.

(Similar or related answers acceptable)

12a. D
12b. D

13a. B
13b. N
13c. B
13d. B
13e. N
13f. N
13g. B
13h. B
13i. B
13j. N
13k. N

14a. B
14b. E
<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15f. A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16a. D</td>
<td>16b. B</td>
<td>16c. C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17a. D</td>
<td>17b. D</td>
<td>17c. C</td>
<td>17d. A</td>
<td>17e. A</td>
</tr>
<tr>
<td></td>
<td>17f. D</td>
<td>17g. C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18a. C</td>
<td>18b. B</td>
<td>18c. C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19. B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20. C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>21a. D</td>
<td>21b. E</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>23a. C</td>
<td>23b. B</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

NOW REFER TO PRESCRIPTION SHEET.
FOR EACH INCORRECT ITEM, IDENTIFY CORRESPONDING LEARNING OBJECTIVE.
THEN REFER TO PRESCRIPTIVE PAGE REFERENCES IN YOUR TEXTBOOK.
Directions: Check your test answers against this Answer Key. Circle the item numbers for every incorrect answer. Then refer to the Prescription Sheet to identify the corresponding learning objectives and the prescriptive page references.

24a. B
24b. A
24c. C
24d. B
24e. A

25a. C
25b. C

26. C

27a. A
27b. A

28. An incorrect response may be as much indicative of failure to comprehend the reading as it is of failure to perform the mathematics correctly.

(Similar or related answers acceptable)

29a. E
29b. C

30a. D
30b. A

31. C
-2-

32a. A
32b. B
32c. C.
32d. C
32e. B
32f. A
32g. B

32h. B
32i. C
32j. A
32k. D
32l. C

32m. A
32n. D
32o. D

33. B

34. D

35a. A
35b. C

36a. D
36b. C
36c. A

NOW REFER TO PRESCRIPTION SHEET.
FOR EACH INCORRECT ITEM, IDENTIFY CORRESPONDING LEARNING OBJECTIVE.
THEN REFER TO PRESCRIPTIVE PAGE REFERENCES IN YOUR TEXTBOOK.
ANSWER KEY

DIAGNOSIS OF MEASUREMENT
COMPETENCY
PART IV--STATISTICAL CONCEPTS

Directions: Check your test answers against this Answer Key. Circle the item numbers for every incorrect answer. Then refer to the Prescription Sheet to identify the corresponding learning objectives and the prescriptive page references.

37a. Rank Test Score  
1 18  
2 19  
3 20  
4 21  
5 28  
6 33  
7 36  
8 37  

37b. Rank Test Score  
1 26  
2 24  
3 19  
4 17  
5 16  

37b. Rank Test Score  
9 38  
10 41  
11 42  
12 44  
13 46  
14 48  
15 50  

38a. B  
38b. C  

39a. A  
39b. B  

40. C  

41a. C  
41b. A  
41c. B  
41d. A  

42a. Mean = 24.4  
Median = 24.5  
Mode = 20  
42b. E
43a. A
43b. A

44a. D
44b. B
44c. B

45. A

46. C

47. $Q_3 = 22 \quad Q_1 = 14.5$  Semi-interquartile range (Q) = 3.75

48a. Percentile rank = 98

48b. | Standard z score | Approximate Percentile Rank |
     |                |                           |
     | -3            | 0.1                       |
     | -2            | 2                         |
     | -1            | 16                        |
     | 0             | 50                        |
     | +1            | 84                        |
     | +2            | 98                        |
     | +3            | 99.9                      |

49a. C

49b. | Area Under Normal Curve | Percentage of Cases In Area |
     |                        |                            |
     | ±1 standard deviation from the mean | 68                          |
     | ±2 standard deviations from the mean | 96                          |
     | ±3 standard deviations from the mean | 99.8                        |

50. C

51. C

52a. D
52b. D
<table>
<thead>
<tr>
<th>53. Standard Score Scale</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>z</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>stanine</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>deviation IQ</td>
<td>100</td>
<td>15</td>
</tr>
<tr>
<td>CEEB</td>
<td>500</td>
<td>100</td>
</tr>
</tbody>
</table>

54. B

55a. T = 36.5

55b. Transformation A Score

<table>
<thead>
<tr>
<th>550</th>
<th>400</th>
<th>600</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Transformation B Score

<table>
<thead>
<tr>
<th>6</th>
<th>3</th>
<th>7</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

56. C

57a. C
57b. B
57c. A

58. D

59a. A coefficient of correlation is a number that tells us to what extent two things are related (i.e., to what extent variations in the one go with variations in the other).

59b. Positive correlation is a relationship where high scores on one variable tend to go with high scores on the other variable, and low with low.

Negative correlation is a relationship where high scores on one variable tend to go with low scores on the other variable, and low with high.

No relationship is a random relationship between two variables.

Perfect relationship is a linear relationship between two variables, where equal increments in one variable correspond to equal increments in the second variable.
60a. E
60b. C
60c. B

61. D
62. B

63. High: From .60 to 1.00
Low: From .00 to .30 (But there are no set rules.)
Moderate: From .30 to .60

64a. D
64b. A
64c. B

65a. A
65b. C
65c. B

66. E
67. A

68a. R
68b. R
68c. I
68d. O
68e. N
68f. O

69a. B
69b. A

70. D

---

NOW REFER TO PRESCRIPTION SHEET.
FOR EACH INCORRECT ITEM, IDENTIFY CORRESPONDING LEARNING OBJECTIVE.
THEN REFER TO PRESCRIPTIVE PAGE REFERENCES IN YOUR TEXTBOOK.
APPENDIX E
# PRESCRIPTION SHEET

## Part I--Standardized Tests

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>Prescription (Stanley &amp; Hopkins, 1972)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Knowledge of advantages and disadvantages of standardized tests.</td>
<td>360-361, 375-376</td>
</tr>
<tr>
<td>2. Ability to compare standardized with teacher-made tests and choose appropriately in a local situation.</td>
<td>360-361, 375-376</td>
</tr>
<tr>
<td>3. Ability to interpret achievement test scores.</td>
<td>365-374, 435-436</td>
</tr>
<tr>
<td>4. Understanding of the importance of adhering strictly to the directions and stated time limits of standardized tests.</td>
<td>148-150, 431</td>
</tr>
<tr>
<td>5. Knowledge of sources of information about standardized tests.</td>
<td>421-423</td>
</tr>
<tr>
<td>7. Knowledge of general information about individual intelligence and aptitude tests.</td>
<td>353-355 (323-358)</td>
</tr>
<tr>
<td>8. Familiarity with need for and application of personality and interest inventories.</td>
<td>380-396</td>
</tr>
<tr>
<td>9. Familiarity with need for and application of projective techniques.</td>
<td>396-402</td>
</tr>
<tr>
<td>10. Knowledge of general uses of tests, such as motivating, emphasizing important teaching objectives in the minds of pupils, providing practice in skill, and guiding learning.</td>
<td>7-11, 418</td>
</tr>
</tbody>
</table>
## Learning Objective

### Prescription

(Stanley & Hopkins, 1972)

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>Prescription</th>
</tr>
</thead>
<tbody>
<tr>
<td>11. Knowledge of advantages and disadvantages of teacher-made tests.</td>
<td>360-361, 375-376</td>
</tr>
<tr>
<td>12. Knowledge of the fact that test items should be constructed in terms of both content and behavior</td>
<td>173-182</td>
</tr>
<tr>
<td>13. Ability to state measurable educational objectives.</td>
<td>172-180</td>
</tr>
<tr>
<td>14. Knowledge of the general principles of test construction (e.g., planning the test, preparing the test and evaluating the test).</td>
<td>182-186</td>
</tr>
<tr>
<td>15. Knowledge of advantages and disadvantages of various types of objective test items</td>
<td>218, 221, 226-227, 236-237, 255, 260-261, 264</td>
</tr>
<tr>
<td>16. Knowledge of the techniques of administering a test.</td>
<td>428-431</td>
</tr>
<tr>
<td>17. Ability to construct different types of test items.</td>
<td>223-226, 230-232, 246-255, 260-262</td>
</tr>
<tr>
<td>18. Understanding and application of correction-for-guessing formula to an objective test.</td>
<td>142-147</td>
</tr>
<tr>
<td>19. Knowledge of the principles involved in scoring subjective and objective tests.</td>
<td>212-214, 431-435</td>
</tr>
<tr>
<td>20. Knowledge of effective procedures in reporting to parents.</td>
<td>312-318</td>
</tr>
<tr>
<td>21. Knowledge of effective marking procedures.</td>
<td>304-312</td>
</tr>
<tr>
<td>22. Knowledge of advantages and disadvantages of essay questions.</td>
<td>197-200, 203-207</td>
</tr>
<tr>
<td>23. Familiarity with the blueprint scheme for dealing with the content and behavior dimensions in test planning.</td>
<td>172-186</td>
</tr>
</tbody>
</table>
## PRESCRIPTION SHEET

Part III--Uses of Measurement and Evaluation

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>Prescription (Stanley &amp; Hopkins, 1972)</th>
</tr>
</thead>
<tbody>
<tr>
<td>24. Ability to interpret diagnostic test results so as to evaluate pupil progress</td>
<td>93-96, 371-375</td>
</tr>
<tr>
<td>25. Ability to interpret the ratio formula relating CA, MA and IQ.</td>
<td>337-339</td>
</tr>
<tr>
<td>26. Familiarity with expected academic behavior of students classified in certain IQ ranges.</td>
<td>346-352</td>
</tr>
<tr>
<td>27. Ability to interpret a profile of sub-test results of standardized tests.</td>
<td>93-96, 371-375</td>
</tr>
<tr>
<td>28. Knowledge of limitations of tests that require reading comprehension.</td>
<td>150-152</td>
</tr>
<tr>
<td>29. Understanding of the limitations of the &quot;percentage&quot; system of marking.</td>
<td>305-306</td>
</tr>
<tr>
<td>30. Understanding of the limitations of applying national norms to a local situation.</td>
<td>80-85, 365-366</td>
</tr>
<tr>
<td>31. Ability to compare two classes on the basis of the means and standard deviations of a test.</td>
<td>48-50</td>
</tr>
<tr>
<td>32. Knowledge of concepts of validity, reliability and item analysis.</td>
<td>110-112, 114-132, 267-280</td>
</tr>
<tr>
<td>33. Ability to do a simple item analysis for a teacher-made test.</td>
<td>267-280</td>
</tr>
<tr>
<td>34. Knowledge of the limitations of ability grouping based on only one measure of ability.</td>
<td>436-437</td>
</tr>
<tr>
<td>35. Knowledge of limitations in interpreting IQ scores.</td>
<td>323-358</td>
</tr>
<tr>
<td>36. Familiarity with the nature and uses of a frequency distribution.</td>
<td>15-23</td>
</tr>
</tbody>
</table>
## PRESCRIPTION SHEET

### Part IV--Statistical Concepts

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>Prescription (Stanley &amp; Hopkins, 1972)</th>
</tr>
</thead>
<tbody>
<tr>
<td>37. Familiarity with techniques of ranking a set of scores.</td>
<td>21-22</td>
</tr>
<tr>
<td>38. Ability to set up class intervals for a frequency distribution.</td>
<td>-----</td>
</tr>
<tr>
<td>39. Understanding of the basic concept of the standard error of measurement.</td>
<td>118-121</td>
</tr>
<tr>
<td>40. Understanding of the nature and uses of the histogram and frequency polygon.</td>
<td>-----</td>
</tr>
<tr>
<td>41. Understanding of the nature and uses of the mode, median, and mean.</td>
<td>16-17, 24-25</td>
</tr>
<tr>
<td>42. Ability to compute the mode, median and mean for simple sets of data.</td>
<td>17-18</td>
</tr>
<tr>
<td>43. Knowledge of advantages and disadvantages of the mode, median and mean.</td>
<td>18-19, 25-28</td>
</tr>
<tr>
<td>44. Understanding of the meaning of the term &quot;variability&quot; and its connection with such terms as &quot;scatter,&quot; &quot;dispersion,&quot; &quot;deviation,&quot; &quot;homogeneity,&quot; and &quot;heterogeneity.&quot;</td>
<td>28-32, 37</td>
</tr>
<tr>
<td>45. Understanding of the nature and uses of the semi-interquartile range.</td>
<td>32-33</td>
</tr>
<tr>
<td>46. Understanding of the nature and uses of the standard deviation.</td>
<td>29-31, 35-37</td>
</tr>
<tr>
<td>47. Ability to compute the semi-interquartile range for simple sets of data.</td>
<td>32-35</td>
</tr>
<tr>
<td>48. Knowledge of the approximate percentile ranks associated with standard scores along the horizontal baseline of the normal curve.</td>
<td>39-40, 47</td>
</tr>
<tr>
<td>49. Knowledge of the percentage of the total number of cases included between + or - 1, 2 or 3 standard deviations from the mean in a normal distribution.</td>
<td>38-40, 47</td>
</tr>
<tr>
<td>Learning Objective</td>
<td>Prescription</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>50. Knowledge of the fact that the normal curve is an ideal distribution, an abstract model approached but never achieved fully in practice.</td>
<td>-----</td>
</tr>
<tr>
<td>51. Knowledge of the limitations of using the normal curve in practice as the fact that in large heterogeneous groups it &quot;fits&quot; most test data rather well and that it aids in the interpretation of test scores, but does not necessarily apply to small selected groups.</td>
<td>-----</td>
</tr>
<tr>
<td>52. Ability to convert a given raw score into a z score from a mean and standard deviation of a set of scores.</td>
<td>43, 48</td>
</tr>
<tr>
<td>53. Knowledge of the means and standard deviations of common standard score scales such as the z, T, stanine, deviation IQ and CEEB scales.</td>
<td>43-44, 46, 48-50</td>
</tr>
<tr>
<td>54. Knowledge of the common applications of standard scores.</td>
<td>42-45, 49-53</td>
</tr>
<tr>
<td>55. Knowledge of how to convert from one type of standard score to another.</td>
<td>44, 46-47, 49-50</td>
</tr>
<tr>
<td>56. Knowledge of the fact that the mode, mean and median coincide for a symmetrical distribution.</td>
<td>18, 38, 48</td>
</tr>
<tr>
<td>57. Knowledge of the meaning of the terms used to designate certain common non-normal distributions such as &quot;positively skewed,&quot; &quot;negatively skewed,&quot; and &quot;bimodal&quot; distributions.</td>
<td>19-20, 49</td>
</tr>
<tr>
<td>58. Knowledge of the fact that any normal distribution can be completely described in terms of its mean and standard deviation.</td>
<td>-----</td>
</tr>
<tr>
<td>59. Ability to define the concept of correlation, including such terms as &quot;positive correlation,&quot; &quot;negative correlation,&quot; &quot;no relationship&quot; and &quot;perfect relationship.&quot;</td>
<td>53-54</td>
</tr>
<tr>
<td>Learning Objective</td>
<td>Prescription (Stanley &amp; Hopkins, 1972)</td>
</tr>
<tr>
<td>-----------------------------------------------------------------------------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>60. Knowledge of the significance of the numerical magnitude and the sign of the Pearson Product-Moment Correlation Coefficient.</td>
<td>53-54, 68-70</td>
</tr>
<tr>
<td>61. Knowledge of the fact that correlation coefficients do not imply causality between two measures.</td>
<td>61-63</td>
</tr>
<tr>
<td>62. Knowledge of the fact that correlation coefficients alone do not indicate any kind of percentage.</td>
<td>54, 61</td>
</tr>
<tr>
<td>63. Understanding of the meaning of a given correlation coefficient in terms of whether it is &quot;high,&quot; &quot;low&quot; or &quot;moderate.&quot;</td>
<td>69</td>
</tr>
<tr>
<td>64. Familiarity with the scatter diagram and the ability to make simple interpretations from it.</td>
<td>60</td>
</tr>
<tr>
<td>65. Knowledge of what size of correlation to expect between two given variables in terms of logical reasoning, e.g., in terms of a common factor.</td>
<td>-----</td>
</tr>
<tr>
<td>66. Understanding of the fact that a raw score has no meaning alone and needs some context in which it can be interpreted.</td>
<td>42</td>
</tr>
<tr>
<td>67. Familiarity with the nature and uses of the common derived scores, viz., age scales, grade scales, percentile scales and standard score scales.</td>
<td>90-92, 366-371</td>
</tr>
<tr>
<td>68. Understanding of certain concepts associated with scale theory, such as types of scales (nominal, ordinal, cardinal and absolute); translation of scores to a common scale; units of equal size; and common reference points (zero or the mean).</td>
<td>-----</td>
</tr>
<tr>
<td>69. Ability to interpret raw scores from a given set of norms.</td>
<td>82-85</td>
</tr>
<tr>
<td>Learning Objective</td>
<td>Prescription</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>70. Understanding of the fact that interpretations of achievement from norms is affected by ability level, cultural background and curricular factors.</td>
<td>(Stanley &amp; Hopkins, 1972) 81, 83-84</td>
</tr>
</tbody>
</table>
INSTRUCTOR'S GUIDE

DIAGNOSIS OF MEASUREMENT COMPETENCY

A Diagnostic-Prescriptive System

For Developing Measurement Competency for Prospective Teachers

by

Allen N. Shub

Loyola University of Chicago
INTRODUCTION

This Instructor's Guide was written to help the Tests and Measurements instructor use the Diagnosis of Measurement Competency materials in his or her course. Diagnosis of Measurement Competency is a diagnostic-prescriptive system for developing measurement competency for prospective teachers. The materials are designed to aid the Tests and Measurements instructor in diagnosing specific skill weaknesses in Tests and Measurements and prescribing appropriate references in the student's textbook for remediation. The system can be used on an ongoing basis throughout the Tests and Measurements course or can be used as a refresher aid for individuals who have already completed such a course.

HOW TO USE

The diagnostic-prescriptive system contains the following components: Diagnostic Tests, Answer Keys, Prescription Sheets, and Survey Tests. The use of each component is described in this section.

Diagnostic Tests

There are four diagnostic tests, entitled Diagnosis of Measurement Competency, corresponding to Mayo's (1967) four content categories of measurement competencies:

Part I--Standardized Tests--covers 10 learning objectives with 31 test items.
Part II -- Construction and Evaluation of Classroom Tests -- covers 13 learning objectives with 44 test items.

Part III -- Uses of Measurement and Evaluation -- covers 13 objectives with 38 test items.

Part IV -- Statistical Concepts -- covers 34 learning objectives with 63 test items.

All four diagnostic tests cover a total of 70 learning objectives with 176 test items.

The four diagnostic tests can be administered in any order. It is suggested, however, that the appropriate diagnostic test be administered following the corresponding instructional unit of the Tests and Measurements course. If desired, furthermore, the student can be tested with only those items that measure the learning objectives of interest to the instructor; in such a case, the instructor would tell the student(s) which items to respond to on the test. Each diagnostic test can be self-administered independently by the student or administered in a group by the instructor.

Answer Keys

Each Diagnosis of Measurement Competency test has a corresponding Answer Key which provides the correct answers for each test item. The student should check his or her work on the test with the Answer Key and,
for reference, circle the item numbers for each incorrect answer. If the instructor desires, he or she can do the scoring for the student. The item numbers on the test correspond to the objective numbers of the system; for example, diagnostic test items 52a and 52b measure learning objective 52--

"Ability to convert a given raw score into a z score from a mean and standard deviation of a set of scores."

**Prescription Sheets**

Each *Diagnosis of Measurement Competency* test and Answer Key have a corresponding Prescription Sheet which provides prescriptive page references in the student's *Tests and Measurements* textbook for each of the learning objectives covered by the diagnostic tests. After the student or the instructor has scored the diagnostic test and circled the item numbers for every incorrect answer, the student or instructor consults the Prescription Sheet. The Prescription Sheet identifies the particular learning objective corresponding to the incorrect items and provides specific page references in the textbook that the student is to consult in order to find material for remediation for the nonmastered learning objectives.

No rule is provided in this Instructor's Guide as to how many items the student must get incorrect before remediation is indicated; that decision is left for instructor's judgment. For those objectives with more than one item to measure them, even one incorrect item may be indicative of a skill deficiency, however, and it is recommended, therefore, that the
student consult the appropriate pages in the textbook for remediation.
(The number of items to measure each learning objective varies, depending on the number of items necessary to properly cover each objective.)

Entry Survey Tests

It is recommended that Mayo's (1967) Measurement Competency Test, Form A or B, be used as an optional Survey Test for those who desire such a component. The Measurement Competency Test can be used as a survey pretest to indicate which of the four diagnostic tests to administer; it can also be used as a general survey posttest. This approach--using the Survey Test--will most likely be used when the system is used for remediation some time after the student has completed a Tests and Measurements course. Thus, the Survey Test is one entry vehicle into the diagnostic-prescriptive system. (See Table A for the item numbers of the Measurement Competency Test which correspond to the four diagnostic tests.)

Other Entry Vehicles

Besides the Survey Test, there are other ways that a student can enter the diagnostic-prescriptive system.

If the system is being used on an ongoing basis throughout the Tests and Measurement course, the appropriate diagnostic test can be administered following the particular unit of instruction, for example, Standardized Tests, in order to determine which of the skills covered in this
<table>
<thead>
<tr>
<th>Diagnostic Test</th>
<th>Learning Objective Numbers</th>
<th>Item Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Form A</td>
</tr>
<tr>
<td>I. Standardized Tests</td>
<td>1-10</td>
<td>1-15</td>
</tr>
<tr>
<td>II. Construction and Evaluation of Classroom Tests</td>
<td>11-23</td>
<td>16-30</td>
</tr>
<tr>
<td>III. Uses of Measurement and Evaluation</td>
<td>24-36</td>
<td>31-45</td>
</tr>
<tr>
<td>IV. Statistical Concepts</td>
<td>37-70</td>
<td>46-60</td>
</tr>
</tbody>
</table>

Note. The data in this table are from Mayo, 1967.
unit have not yet been mastered by the students. Or, the appropriate diagnostic test can be administered prior to the instructional unit in order to assist the instructor in concentrating his or her instruction to the areas most needed by the students.

Another entry into the system is through the instructor’s observations. The instructor may observe that certain students are having difficulty with, say, Statistical Concepts and elect to administer that diagnostic test to those students in order to help determine the specific weaknesses in that area.

Other uses of the system will no doubt become apparent to the instructor through continued use of it. The instructor is encouraged to be innovative; the system was designed to be flexible enough to meet most classroom needs.

Recordkeeping

If the students use the diagnostic-prescriptive system independently, the instructor need not maintain records of the students' progress through the system unless desired. For those instructors who desire to track their students, a Progress Chart may be prepared. One suggestion is a chart with 70 columns to represent the 70 learning objectives and with as many rows as there are students in the class. The instructor can mark a slash in each cell to indicate that remediation for that objective is indicated and an X
in that cell to indicate that remediation has been completed. Other methods of recordkeeping may be more convenient for the instructor.

REFERENCE

The dissertation submitted by Allen N. Shub has been read and approved by the following committee:

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

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

Dr. Ernest I. Proulx
Professor, Curriculum & Instruction and
Guidance & Counseling, Loyola

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

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

5/7/77
Date

Samuel T. Mayo
Director's Signature