8-1974

A Study of the Relationship between Norm-Referenced Tests and Criterion-Referenced Tests

Marilyn W. Van Valkenburgh
Western Michigan University

A STUDY OF THE RELATIONSHIP BETWEEN NORM-REFERENCED TESTS AND CRITERION-REFERENCED TESTS

by

Marilyn W. Van Valkenburgh

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment
of the
Degree of Doctor of Education

Western Michigan University
Kalamazoo, Michigan
August 1974

ACKNOWLEDGEMENTS

With appreciation the author wishes to acknowledge Drs. Rodney W. Roth and Bradley E. Huitema, members of the committee, for their expertise in research matters related to the investigation. My advisor, Dr. Donald C. Weaver, deserves special recognition for his constructive suggestions and continued support throughout the completion of this study and doctoral program.

In addition, the Kalamazoo Public Schools provided the data, and many hours of assistance were contributed by their competent staff in the Department of Research under the direction of Dr. Charles Townsend. Mr. Mark Rennhack was very knowledgeable in the use of computers, while Dr. David Bartz, through constant support and continued interest, made the operational aspects of the study particularly efficient.

Whereas the continued interest and support of my parents made the doctoral studies possible, the day-by-day assistance and encouragement from my husband, Bill, and our three children, Paula, Katrina, and William, made the completion of the dissertation and doctoral program a reality.

Marilyn W. Van Valkenburgh

74-28,853

VAN VALKENBURGH, Marilyn W., 1929-
A STUDY OF THE RELATIONSHIP BETWEEN NORM-REFERENCED
TESTS AND CRITERION-REFERENCED TESTS.

Western Michigan University, Ed.D., 1974
Education, general

Xerox University Microfilms, Ann Arbor, Michigan 48106

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

STATEMENT OF THE PROBLEM

Introduction

Before the advent of criterion-referenced tests, decisions made by the educational leader regarding a testing program and its relationship to the total curriculum were based upon norm-referenced tests which have been standard fare in our educational institutions for many years. With the addition of criterion-referenced tests, decisions regarding tests have become more complex due to the existing uncertainty in many situations as to whether criterion-referenced tests should be utilized rather than the traditional norm-referenced tests.

Proponents of criterion-referenced tests, in particular, have led "campaigns" to initiate criterion-referenced testing on the local, state, and national levels in programs of student assessment. Such promotional campaigns imply that it is possible, through the use of criterion-referenced tests alone, to provide the necessary information to make valid educational decisions. This situation has led to considerable conflict, as evidenced by recent debates at all levels of the educational/governmental scene.

On the national level on March 22, 1973, as reported in the Congressional Record, Maryland's Senator Beall, who is a member of the Committee on Labor and Public Welfare, stated:

> There is a need to improve our techniques for testing reading ability and achievement. There is already some interesting work going on as evidenced by the Educational Commission on the States' national assessment of educational processes, and also the work in my State on criterion-referenced tests [p. S 5373].

In a report dated August 15, 1973 from the Committee on Education and Labor to "Education Writers, Education Associations, Chief State School Officers, and School Board Members" concerning the Revised Elementary and Secondary Educational Act, an amendment was offered providing for Title I funds to be distributed according to educational deprivation as opposed to the traditional poverty criteria. The report indicated that Minnesota's Representative Quie had introduced a bill which would identify educational deprivation through the use of criterion-referenced tests. It further stated:

> Under that proposal, a national sample test would identify concentration of educationally deprived children between the States, and Title I money would follow. Then the States would have their own assessments to distribute the money to the local schools. Quie insisted on the use of criterion-referenced tests, which measure a child's attainment of specific goals in reading and math, instead of standardized norm-referenced tests, which compare achievement between children [p. 1].

The subcommittee did not buy the complete bill, but several of the elements were incorporated in the ESEA amendments, as indicated:

> A national Commission was approved to find out whether it is possible to distribute Title I (monies) based on pupil performance on criterion-referenced tests. The commission, if enacted, will report back March 31, 1976 [p. 1].

On the state level, Michigan State Superintendent of Public Instruction Porter has indicated that if ESEA Title I funds can be distributed according to the educational needs of students instead of poverty levels, Michigan will be one of the first states to do so. As of 1973, Michigan had already turned to criterion-referenced tests for its state assessment testing program.

As has been indicated, there is a great amount of debate at all levels of the educational/governmental scene concerning the use of criterion-referenced and norm-referenced tests. Unfortunately, little research information concerning criterion-referenced tests is available at this time to aid educational and/or governmental leaders in making vital testing decisions regarding an effective testing program and its relationship to instructional planning.

The Problem

The purpose of this study was to determine what information can be provided by criterion-referenced tests to aid the educational leader in making instructional decisions.

Specifically, this investigation determined whether or not information provided by the results of a criterion-referenced test can predict relative performance, i.e., approximate grade equivalent, as indicated on a norm-referenced test.

For purposes of clarity, the problem statement is divided into three separate discussions of the following problems related to the purpose of the study: (1) norm-referenced and criterion-referenced tests, (2) decision-making and testing, and (3) interpretation of test results.

Problems of norm-referenced and criterion-referenced tests

Aside from the national and state political/educational testing debates, the educational leader is faced with many decisions concerning the use of tests. Norm-referenced tests provide information as to how a particular person and/or group of persons performs in relation to a comparable population. Whereas information provided by norm-referenced tests is useful, a need continues to exist for additional kinds of descriptive information. Knowing how an individual or group compares with other individuals or groups is informative, but provides little information as to what specific skills have been mastered in a particular subject or area.

Thus, with the current emphasis on goals and specific objectives relating to various subjects, individualization via instructional systems (such as Individually Prescribed Instruction), and taxonomies of learning, criterion-referenced tests have become a reality. Through the use of criterion-referenced tests, it is now possible to diagnose the strengths and weaknesses of individuals and classes regarding specific skills needed to be mastered; but comparing the relative standing of an individual and/or class with a comparable population assumes the use of data produced through the administration of a norm-referenced test.

Problems concerning decision-making and testing

Garvin (1970) states that the ultimate purpose of measurement is "to inform decision-making" which, of course, would aid the educational leader in making the best possible decisions. Many of these decisions relate to the instructional process itself. In fact, Katz (1972) notes that one general purpose of testing is "to improve instruction," involving five "intermediate" purposes of testing as follows: (1) placement, (2) diagnosis, (3) assessment, (4) prediction, and (5) evaluation (p. 176). In an attempt to further clarify the functions of the above-stated "intermediate" purposes of testing as cited by Katz, each will be discussed.

Placement refers to placing students in relation to one another in various groups (selection) and, also, to placing a student at an appropriate level in an instructional sequence of content in a subject. For example:

1. Placing students in particular classes, sections, reading groups, instructional groups, and remedial groups.

2. Placing a student in materials and methods appropriate to "where he is" in regard to his own skills and development in a particular subject (content) area, irrespective of where his fellow students may be functioning.

As can be noted, placement may include both relative comparisons of persons (inter-person comparisons) and/or a comparison of one's own development with respect to the content of a subject (intra-person comparison). Also, such placements may be made within a classroom as well as within a school.

Diagnosis involves analyzing in depth the strengths and weaknesses of particular students in regard to their skills, knowledge, and style of learning. The term can also apply to diagnosing the needs of a whole class and the individual needs within a class.

Assessment involves measuring the effectiveness of a teaching method or treatment, and can be utilized to indicate the amount of student growth and development.

Prediction utilizes measurement for the purpose of forecasting an individual's future performance on the basis of test results.

Evaluation (according to Katz) involves the use of tests to compare one school with other comparable schools. It indicates a broader usage of test results than the four purposes mentioned previously.

According to Katz (1972), all of the above purposes involve comparisons of some kind. The various comparisons require various standards. Therefore, criterion-referenced tests would be appropriate for some of the above purposes, while norm-referenced tests would be more suitable for others. Currently, it is necessary to use norm-referenced tests to obtain information for decisions relating to placement, diagnosis, assessment, prediction, and evaluation. Criterion-referenced tests are necessary, however, and are currently being used to obtain specific information needed in decisions involving placement (of the student in materials and methods appropriate to his own skills and development in a particular subject), diagnosis, assessment, and, to a limited extent, evaluation.

Problems concerning interpreting test results

As the situation exists today, the educational leader of the 1970's will have to be thoroughly familiar with both criterion- and norm-referenced tests so decisions can be made as to *when* each type should be used and for what *purposes*. In addition, persons in education will need to become sophisticated in the interpretation of criterion-referenced tests, as evidenced by the following problem:

> On the fourth-grade 1973-74 Michigan Educational Assessment Program (criterion-referenced

test), a student missed the following reading objective, which is not exclusively a fourth-grade skill: "Indicate author's purpose." This is all the information available, so one must ask several questions concerning:

1. Test results
    a. Is the student really unable to identify the author's purpose?
    b. Was the material too difficult for him/her to read and, therefore, it was missed?
    c. Was the student not feeling well, not focusing on the task, etc.?
2. Implications for instruction
    a. What materials are appropriate for instruction?
    b. Should the student continue using fourth-grade reading materials and work on identifying the author's purpose?
    c. Should the student be given easier materials and then determine if the student has particular difficulty with this specific skill?

Thus, with the use of criterion-referenced tests, one has the advantage of having results reported by specific skills but lacks the grade level scores which could be useful in the interpretation of the test results <u>and</u> in supplying the appropriate level of materials to the student.

Nitko (1970) states that "under certain circumstances <u>both</u> criterion-referenced information and norm-referenced information are needed to make a broad interpretation of an individual's test performance [p. 8]." Other authors (Ebel, 1970; Flanagan, 1951; Glaser, 1973) also note this, but this would result in the necessity of giving both a criterion-referenced and a norm-referenced test. In the light of current criticism concerning the "overtesting" of students and lack of financial resources in education, however, it would seem inadvisable and highly unrealistic to suggest that both a norm-referenced test and a criterion-referenced test be given to groups of students in order to obtain all the information needed for effec'ive instructional planning and placement.

Therefore, if criterion-referenced tests could also report how a particular person and/or class or group of persons performed in relation to a comparable population as in norm-referenced tests, would it not be possible to base most instructional decisions on information provided by criterion-referenced tests? The purpose of the study was to determine if such a possibility exists. If possible, testing could be far more effective and efficient than at present, and criterion-referenced tests alone could be utilized in decisions of instructional management involving placement, diagnosis, assessment, prediction, and evaluation. Thus, if norm-referenced test information could be predicted from a criterion-referenced test, it would have extensive utility and value in the instructional decision-making process.

Questions

According to Frymier (1972), Combs has stated that "measuring what we know how to measure is no substitute for measuring what we need to measure [p. v]." If testing is going to be one of our professional means as opposed to being "legitimized as educational ends [p. v]," then a constant search for the improvement of testing concepts and procedures must continue. Thus, this study was an initial endeavor in that direction in which the relationship between criterion-referenced tests and norm-referenced tests was explored. More specifically, the purposes of this study were to investigate the following questions:

1. What is the relationship between norm-referenced tests and criterion-referenced tests with respect to predicting student performance scores or grade equivalents as indicated on a norm-referenced test?

2. What is the relationship between scores on norm-referenced tests and criterion-referenced tests for (a) fourth- and seventh-grade students, (b) black students and white students, and (c) male and female students?

3. What information can the criterion-referenced test provide educational decision makers in decisions pertaining to placement, diagnosis, assessment, prediction, and evaluation?

Definition of Terms

The following definitions were used to specifically delineate terms used throughout the study:

1. Criterion-referenced test: A test which provides information in terms of specific behaviors mastered by an examinee without reference to the performance of other pupils.

2. Norm-referenced test: A test which provides information in terms of a person's relative standing in relation to an identifiable norm group.

Organization of the Report

The purpose of Chapter I has been to state the problem and its significance, present the questions for investigation, define essential terms, and outline the organization of the report for the reader.

Chapter II includes discussions of the history and development of criterion-referenced tests, characteristics of criterion-referenced tests, characteristics of norm-referenced tests, distinctions between norm- and criterion-referenced tests, and the utility of grade equivalent scores.

Chapter III consists of a brief review of the problem, the population and sample, instrumentation, procedures, and method of data analysis used in the study.

Chapter IV contains a description and analysis of data for specific questions posed for investigation.

Chapter V concludes the study with a summary of the study and presentation of conclusions, implications, and recommendations.

CHAPTER II

REVIEW OF THE LITERATURE

The literature related to the problem identified in Chapter I is divided into five sections: (1) the history and development of criterion-referenced tests, (2) characteristics of criterion-referenced tests, (3) characteristics of norm-referenced tests, (4) distinctions between norm- and criterion-referenced tests, and (5) utility of grade equivalent scores.

Although the history of the first written tests are attributed to the Chinese as early as 1000 B.C., in America the first recorded instance of written examinations was in 1845 in Boston. They were given with the intention of refuting the charges of Horace Mann, Secretary of the Massachusetts State Board of Education, that academic weaknesses existed in the schools. The tests "proved" that Mann's charges were justified (Nunnally, 1972, pp. 13-14).

In the twentieth century, as the development in measurement in psychology occurred and tests were devised to help in the screening and classification of persons in the military during World War II, educational testing also developed. In 1923, the Stanford Achievement Test--which was the first standardized achievement battery--was published, and revisions of that test are still being used today (Nunnally, 1972, p. 17).

History and Development of Criterion-Referenced Tests

Criterion-referenced measures have been in existence for many years and are not necessarily a new development, but rather a redevelopment of past philosophies and measures as evidenced by the following statement by Ebel (1970):

> Contrary to the impression that exists in some quarters, criterion-referenced measurements are not a recent development that modern technology has made possible and that effective education requires [p. 8].

Davis (1971) also acknowledges that since "time immemorial" teachers have been measuring the level of student performance on materials and processes taught by using tests which measure a student's performance against a criterion performance; he substantiates the fact by noting the following instance:

> In 1864 . . . Chadwick wrote that the Reverend George Fisher had prepared a book called the Scale Book, "which contains the numbers assigned to each degree of proficiency in the various subjects of examination . . . The numerical values for spelling . . . are made to depend upon the percentage of mistakes in writing from dictation sentences from works selected for the purpose, examples of which are contained in the 'Scale Book' in order to preserve the same standard of difficulty" [p. 4].

It is of interest to note a very early discussion by Thorndike (1913), in The Original Nature of Man, Educational Psychology, concerning some of his notions relating to the use of criterion-referenced measures. Thorndike wrote:

> During the last thirty years there has been a very strong movement from detailed to crude records of achievement, and from publicity to secrecy . . . . The reasons alleged for the change have been that detailed grades and publicity encourage a pupil to work for "marks," and for excellences in the sense of excelling others, instead of for knowledge or power, and for excellence in the sense of improvement . . . . In my opinion the change was an extremely wasteful way of avoiding one evil by the unnecessary sacrifice of all its attendant goods.
>
> School marks functioned as a measure of superiority and inferiority amongst pupils, and of little else. A pupil who made excellence an aim of his school work was encouraged by every feature of the school's measurements of his work to think of excellence as excelling others--relative achievement--outdoing someone else. Finding that pupils did so, and being rightly suspicious of this gross form of emulation as an end in education, school officers took the easy, but wasteful, way of depriving the pupil of any save the vaguest knowledge of his achievement. To keep him from focussing his attention upon his achievement in comparison with his fellow students' achievements, they kept from him any detailed record whatsoever of his achievement.
>
> To work for marks is not intrinsically bad. If the marks are, as they should be, correct measures of either the amount of knowledge, power, appreciation and skill attained or the amount of progress made, to work for marks means simply to work for knowledge, power, increase in knowledge and power and the like as recognized and measured. The detailed nature and the report to the individual of his school marks were not vices of the old system. Its vice was its relativity and indefiniteness--the fact already described that a given mark did not mean any defined amount of knowledge, or power, or skill--so that it was bound to be used for relative achievement only.

> Suppose, for example, that instead of the traditional "89"s or "good"s a pupil had records of just how many ten-digit additions he could compute correctly in five minutes, of just how difficult a passage he could translate correctly at sight, and of how long it required, and the like. He could, of course, still compare himself with others, but he would not be compelled to do so. He could be encouraged, instead, to compare his present achievement with last month's, to beat his record, or the record for an average ten-year-old . . . .
>
> Rivalry with one's own past and with a "bogey," or accepted standard, is entirely feasible, once we have absolute scales for educational achievement comparable to the scales for the speed at which one can run or the height to which one can jump. Such scales are being constructed [pp. 286-287].

In this discussion Thorndike referred to a classroom experiment in which Kirby found that, by reporting to each fifth-grade student his absolute achievement on "measured tests in addition" and then giving the student 60 minutes of drill, the results were "an improvement of over 50 percent in speed with a slight gain in accuracy as well [p. 288]." In describing the method used, Kirby (1913) noted:

> The children were told that their individual improvement was to be measured and they were shown that no matter how low or high their present record their final standing would be determined by the amount of gain made. They were shown that it was not primarily a contest among individuals of the class, but an effort on the part of each one to surpass his own previous record. The children were encouraged to compare their last record with their own previous records, and at times the scores were read to them in such a way as to indicate gains made. A hypothetical curve was drawn upon the board to indicate the ascent that

> would result from supposed gains, as well as to show them how to keep their own individual curves [p. 8].

Approximately a decade after Thorndike of Columbia University wrote his notions concerning criterion-referenced measures, Morrison of the University of Chicago developed his method of teaching based on mastery learning. Ebel (1970), in discussing Morrison's method of teaching, describes it as follows:

> Morrison . . . developed and popularized a method of teaching based on the mastery of "adaptations" of understanding, appreciation or ability. These, unlike skills, seemed to Professor Morrison not to be matters of degree. " . . . the pupil has either attained it or he has not." To achieve such an adaptation the instructor should organize his materials into units, each focused on a particular adaptation. He should then follow a systematic teaching routine: teach, test, reteach, retest, to the point of actual mastery [p. 6].

Ebel notes that Morrison's ideas had support for some time through 1930, but then became little known or practiced.

Besides Morrison's Unit-Mastery Plan, two other plans The Winnetka Plan and the Dalton Plan, also enjoyed popularity for a time. The Winnetka Plan required frequent testing of students to insure the mastery of specified skills at a predetermined level, whereas the Dalton Plan required the student to sign a contract stating that the student would reach certain competencies and levels of performance before advancing to the next unit. In both cases, the level of performance was measured by mastery tests.

More recently, a revival of similar philosophies and mastery testing within the educational setting has been credited to Glaser, by Nitko (1970). The origin of the term "criterion-referenced tests" has been attributed to Glaser "in connection with proficiency measurement in training . . . and later was applied to the measurement of educational achievement . . . [Nitko, 1970, p. 2]." Nitko further states that "the motivation for this application of achievement measurement stemmed from a concern about the kind of achievement information required to make instructional decisions [p. 2]."

In the past several years criterion-referenced tests have been designed to meet the measurement needs of the new instructional models which are based on specific performance objectives. Some of these models include Glaser's Individualized Instruction, Flanagan's Project PLAN, and Carroll's Model of School Learning (Hambleton and Gorth, 1970, p. 3). Currently, criterion-referenced tests are also being used in state assessment programs (Michigan), and their use is being considered for national assessment programs.

Characteristics of Criterion-Referenced Tests

> A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards [Glaser and Nitko, 1971, p. 653].

In the 1960's, Glaser introduced the term "criterion-referenced test" in an effort to identify a "new" type of test and to make the needed distinction between it and the traditional norm-referenced test which had been used in our country since the 1920's. While the norm-referenced test was designed to measure and compare an individual's performance against the performance of others in a similar population, the criterion-referenced test was designed to measure an individual's performance against a predetermined set of clearly specified standards. Such specific performance standards represent "a class or domain of tasks" which are considered to be essential for an individual to master (Glaser and Nitko, 1971).

In a further clarification of the nature of criterion-referenced tests, Nitko (1970) has reported the following four characteristics to be essential and inherent in criterion-referenced tests:

1. The classes of behaviors that define different achievement levels are specified as clearly as is possible before the test is constructed.

2. Each behavior class is defined by a set of test situations (that is, test items or test tasks) in which the behaviors can be displayed in terms of all their important nuances.

3. Given that the classes of behavior have been specified and that the test situations have been defined, a representative sampling plan is designed and used to select the test tasks that will appear on any form of the test.

> 4. The obtained score must be capable of expressing objectively and meaningfully the individual's performance characteristics in these classes of behavior [p. 8].

Defining a domain of tasks and related problems: Various viewpoints

The first characteristic named by Nitko indicates that the initial consideration in criterion-referenced test development is to define a class or domain of tasks which will constitute a set of clearly specified standards, the mastery of which is considered to be essential to the learner's progress. On the surface, developing a set of clearly specified standards appears to be a reasonable and manageable task; however, it has been the subject of discussion of many authors with varying viewpoints.

Glaser (1973) has noted that "the standard (or criterion) against which a student's performance is compared . . . is the behavior which defines each point along the achievement continuum [p. 7]." Each point along the continuum must be very specific and explicit, forming a set of behaviorally stated objectives which serve as the basis for criterion-referenced test construction.

On the other hand, Roudabush and Green (1971) have raised concerns in noting that the use of a criterion-referenced test "presupposes a specific knowledge of terminal behaviors [p. 1]" which are considered to be desirable; it also "presupposes that this inventory of behaviors will be comprehensive for the domain being measured [p. 1]." Accordingly, Roudabush and Green have stated that the selection of a set of objectives or behaviors to be sampled by a criterion-referenced test is based on subjective judgments. The authors note that, even in a discipline as well structured as mathematics, "the judgments of the value of the objectives to be measured by the tests are both difficult and critical [p. 2]."

In contrast to Glaser's view and similar to the viewpoints of Roudabush and Green, Cartier (1968) questions the assumption that "a complete and unambiguous inventory can be made of all the behaviors necessary for adequate performance [p. 29]." Furthermore, if this is possible, then an inventory of the most essential skills may result in several thousand individual behavioral objectives which pose additional problems for test construction.

Similarly, this problem has also been cited by Roudabush and Green (1971). If a criterion-referenced test is to be "comprehensive for a discipline and if it covers material generally taught over a period of years [p. 3]," the number of objectives needed becomes <u>very</u> large. Thus, the test becomes extremely long, and/or very few items can be devoted to a specific behavior. The number of test items measuring each objective becomes a crucial problem if the test is to be reliable and provide accurate and specific information concerning an individual (p. 3).

Perhaps Gronlund (1973) has aptly summarized the situation and related problems in the following discussion. Gronlund has suggested that criterion-referenced tests can be used both for measuring mastery of minimum essentials and for measuring development beyond the minimum level. In mastery learning, the domain to be measured is more limited and can be more clearly defined and specified. Beyond mastery level, "the infinite number of available learning tasks and the increased complexity of the learning outcomes pose problems that can be dealt with in only an approximate and tentative manner [p. 7]."

Gronlund (1973) has also offered principles to provide an operational framework relating to criterion-referenced tests. Criterion-referenced testing requires:

1. A clearly defined and delimited domain of learning tasks be identified.
2. Instructional objectives be clearly defined in behavioral (performance) terms.
3. Standards of performance be clearly specified.
4. Student performance be adequately sampled within each area of performance.
5. Test items be selected on the basis of how well they reflect the behavior specified in the instructional objectives.
6. Scoring and reporting system that adequately describes student performance on clearly defined learning tasks.

Problems of criterion-referenced tests: Various viewpoints

While Gronlund's list of principles is very similar to Nitko's list of characteristics of criterion-referenced tests, many authors have cited numerous problems relating in general to criterion-referenced tests. Gronlund (1973) has suggested that there may be problems with most of the above requirements which he stated for criterion-referenced testing. More specifically, Klein (1970) has suggested that criterion-referenced testing would be

> a laudable practice if one knew how to determine what criterion objectives to specify, or what level of performance constitutes their attainment, or how to interpret the results if the objectives are or are not achieved [p. 3].

Roudabush and Green (1971) emphasize not only the role of subjective judgments in determining if an objective (behavior) has value, but also what number of correct solutions represent mastery, and in determining if the test items adequately "represent the desired behavior domain [p. 2]."

Use of criterion-referenced tests in instruction: Various viewpoints

Hambleton and Gorth (1970) have suggested that criterion-referenced tests can be used to serve two purposes: (1) to provide information pertaining to the performance levels of individuals and (2) to evaluate the effectiveness of instruction (p. 4). In providing very specific information as to which objectives are mastered and/or not mastered on a criterion-referenced test by an individual, effective instructional planning to meet the specific needs of an individual can be facilitated.

Roudabush (1973) has reported that, if the specific information provided has utility, a criterion-referenced test should:

> (1) accurately reflect the examinee's standing with respect to the curriculum; that is, show his specific strengths and weaknesses, (2) accurately reflect changes when the examinee's capability to perform has changed, and (3) lead to appropriate decisions for the further instruction of the examinee [p. 2].

In evaluating the effectiveness of instruction, Hambleton and Gorth (1970) suggest using criterion-referenced tests, "combined possibly with the notion of item-examinee sampling [p. 4]," in order to obtain specific results related to the instructional objectives.

Gronlund (1973) has suggested that criterion-referenced tests are useful in teaching in the following ways:

> Formative tests are used during the instructional process to stimulate, build, and evaluate student learning and to appraise the ongoing effectiveness of the instructional procedures. Summative tests are used at the end of instruction to determine what students have learned [p. 6].

Several authors (Gronlund, 1973; Nitko, 1970) have suggested that criterion-referenced tests have been employed in, and perhaps are best fitted to, situations in which the notion of mastery learning is advocated.

In conclusion, Gronlund (1973) further explains that criterion-referenced tests are most useful in classroom instruction where the "learning outcomes are relatively simple and mastery provides a realistic standard of performance [p. 22]." Gronlund further states, however, that "when the instructional outcomes are complex and the purpose is for the student to achieve maximum level of performance [p. 22]," then a norm-referenced test may be more appropriate.

Characteristics of Norm-Referenced Tests

> Norm-referenced measures are those which are used to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device. The meaningfulness of the individual score emerges from the comparison. It is because the individual is compared with some normative group that such measures are described as norm-referenced [Popham and Husek, 1969, p. 19].

Unlike criterion-referenced tests, norm-referenced tests have been widely accepted in the educational setting for many years. Most ability and achievement tests administered in the schools are norm-referenced, as they are scored by comparing an individual's performance with the performance of a comparable group of subjects.

Norms are usually established through a standardization process "in which the procedure, apparatus, and scoring have been fixed so that precisely the same testing procedures can be followed at different times and places [Cronbach, 1970, p. 27]." Though a standardized norm-referenced test may require years to construct and standardize, the explicit procedures have been developed over the last 70 years and are well established, such that they are routinely accepted by testing specialists.

The norm-referenced test measures content and behaviors common to the specific population for which the test was written. Most norm-referenced tests focus on elements of instructional areas which are common to most schools, thus allowing the test to be given to persons at different times and at different places, as the directions for administering and scoring the tests are very specific and controlled.

The norms given on the test allow the comparison of an individual's level of performance with the levels of performance of a comparable population. Therefore, a particular score on a norm-referenced test is compared to some relevant norm distribution, and the test is constructed in such a way as to maximize the variability of test scores. In so doing, the "test is likely to produce fewer errors in ordering individuals on the measured ability [Hambleton and Gorth, 1970, p. 5]."

Besides having standard content and specified norms, the norm-referenced test consists of carefully selected items which are designed to measure the specific trait or sample of behavior desired. In order to maximize test score variance, items which are too easy or too difficult are eliminated, since they would not produce the desired variance. Therefore, rather than the items being selected because they relate to certain specific objectives, items are chosen by their discriminating power to "spread people out" over a continuum. Also, items which do not seem to measure the same trait as the majority of the other items in the test will be removed. Thus, in a norm-referenced test the items must have a certain level of difficulty and discriminating power in order to produce the kind of variance needed for a norm distribution. As has been mentioned, the process of constructing a norm-referenced test is based on well-defined classical test theory such that item selection, validity, and reliability requirements are clearly specified.

The norm-referenced test may have several equivalent forms, which measure the same sample of behaviors but use different items which have been carefully selected. Therefore, the forms can be used interchangeably. Norm-referenced tests may have comparable forms which span several grades, i.e., grades 1-3, 4-6, 7-9, and 10-12. Comparable tests are frequently used in a school system to insure continuity in testing.

Gronlund (1971) has summarized the characteristics of norm-referenced tests as follows:

> 1. The test items are of high technical quality. They have been developed by educational and test specialists; tried out

> experimentally (pretested); and selected on the basis of difficulty, discriminating power, and relationship to a clearly defined and rigid set of specifications.
>
> 2. Directions for administering and scoring are so precisely stated that the procedures are standard for different users of the test.
>
> 3. Norms, based on representative groups of individuals, are provided as aids in interpreting the test scores. These norms are based on various age and grade groups on a national, regional, or state level. Norms for special groups, such as private schools, might also be supplied.
>
> 4. Equivalent and comparable forms of the test are typically provided as well as information concerning the degree to which the forms are comparable.
>
> 5. A test manual and other accessory materials are provided as guides for administering and scoring the test, for evaluating its technical qualities, and for interpreting and using the results [p. 270].

Gronlund (1971) also notes that, even though norm-referenced tests have common characteristics, there are great differences in the behaviors sampled and in the purposes for which they are used. Because of the above characteristics of norm-referenced tests, Gronlund (1971) notes their usefulness in the following instructional purposes:

> 1. Evaluating the general educational development of pupils in the basic skills and in those learning outcomes common to many courses of study.
>
> 2. Evaluating pupil progress during the school year or over a period of years.
>
> 3. Grouping pupils for instructional purposes.

4. Diagnosing relative strengths and weaknesses of pupils in terms of broad subject or skill areas.

5. Comparing a pupil's general level of achievement with his scholastic aptitude [p. 271].

Roudabush (1973) concludes that a norm-referenced test should:

1. Accurately reflect the examinee's standing with respect to the norm group, that is, show his relative position on the underlying quantity or trait being measured.

2. Accurately predict what the examinee will be able to do successfully [p. 2].

Distinctions Between Norm-Referenced and Criterion-Referenced Tests

Many distinctions between norm-referenced tests and criterion-referenced tests can be noted. In this study, the distinctions of each type of test are listed below and include the following areas of comparison: scores, primary purpose, test development, test items, reliability, validity, reporting and interpretation, uses, and limitations. (Those comments without footnotes are credited to Popham and Husek, whose article is included in Popham's 1973 publication, Criterion-Referenced Measurement. All other comments have the appropriate sources cited.)

I. Scores

A. Norm-referenced test

1. Relative.

2. Meaningfulness of individual's score is dependent upon comparison with some normative group, i.e., dependent on relative position of the score in comparison with other scores (Roudabush, 1973, p. 2).

3. Provides scores in percentiles, stanines, and grade equivalents (Cartier, 1968, p. 29).

B. Criterion-referenced test

1. Absolute.

2. Meaningfulness of individual score is dependent upon comparison with a performance standard or established criterion from curriculum represented, rather than other individuals (Roudabush, 1973, p. 2).

3. Provides score for each objective as mastered or not mastered (Cartier, 1968, p. 29).

II. Primary Purpose

A. Norm-referenced test

1. Makes decisions about individuals, particularly in situations requiring selection.

2. Will discriminate well between examinees who have differing amounts of achievement in the general area of interest (Roudabush, 1973, p. 2).

3. Generally intended to be descriptive and predictive and will accurately predict what the examinee will be able to do successfully (Roudabush, 1973, p. 2).

B. Criterion-referenced test

1. Makes decisions about both individuals and treatments, e.g., instructional programs.

2. Will discriminate well between mastery and nonmastery of the objectives making up the curriculum of interest (Roudabush, 1973, p. 2).

3. Generally intended to be diagnostic and prescriptive and can lead to appropriate decisions for further instruction of the examinee (Roudabush, 1973, p. 2).

III. Test Development

A. Norm-referenced test

1. Samples course objectives (Cartier, 1968, p. 28).

2. Since norm-referenced tests are constructed with the purpose of setting persons apart, the more variability, the better.

B. Criterion-referenced test

1. Tests every essential behavior (Cartier, 1968, p. 28).

2. Variability is not a necessary condition for a good criterion-referenced test. Since current treatments of validity, reliability, and formulas for item analysis are all based on desirability of variability of test scores, does not apply to criterion-referenced test construction.

IV. Test Items

A. Norm-referenced test

1. The item-writer writes items for the purpose of producing variability in the scores, thus eliminating items

which are "too easy" or "too difficult." Attempts to increase the "allure" of wrong answer options.

2. If item is missed by many persons, the item is revised (Cartier, 1968, p. 29).

3. Items should be sensitive to individual differences (Roudabush, 1973, p. 2).

4. Use of item analysis procedures (discrimination indices) to identify items which do not properly discriminate between individuals. Usually they are items which are "too easy," "too difficult," and/or ambiguous.

B. Criterion-referenced test

1. The item-writer makes certain each item is an accurate reflection of the criterion behavior.

2. If item is missed by many persons, then the course may be revised (Cartier, 1968, p. 29).

3. Items should be sensitive to instruction (Roudabush, 1973, p. 2).

4. Use of discrimination indices must be modified.

V. Reliability

A. Norm-referenced test

1. Can apply classical procedures because they are dependent on score variability to get estimate of reliability.

B. Criterion-referenced test

1. Must be internally consistent, but classical test procedures for measuring reliability are not appropriate for criterion-referenced tests.

Criterion-referenced tests could be highly consistent, either internally or temporally, and yet indices are dependent on variability and might not reflect that consistency.

VI. Validity

A. Norm-referenced test

1. Can apply classical procedures because they are dependent on score variability to get estimate of validity.

B. Criterion-referenced test

1. Content validity which involves a carefully made judgment, "based on the test apparent relevance to the behaviors legitimately inferable from those delimited by the criterion," is usually employed.

VII. Reporting and Interpretation

A. Norm-referenced test

1. For reporting on individuals, uses group-relative description such as percentile rankings, standard scores, or grade equivalent scores.

2. For reporting on treatments, Popham and Husek consider norm-referenced test to be less than suitable device for this purpose, since emphasis is on producing heterogeneous performance rather than on reflecting treatments or objectives.

B. Criterion-referenced test

1. For reporting on individuals, one can indicate a proficiency level such as 90-percent minimum (objective has been achieved).

2. For reporting treatment assessment, report (a) number of persons who

achieved an established criterion and (b) descriptive statistics such as means and deviations.

VIII. Uses

A. Norm-referenced test

1. Use in instructional sequences where there are several different sequences differing widely in rigor (Garvin, 1973, p. 62).

2. Use of individual selection where degree of selectivity is required because of constraint on the number of individuals who can be admitted (Glaser and Nitko, 1971, p. 655).

3. Provides information for evaluating merits of instructional program.

B. Criterion-referenced test

1. Use in instructional sequence where content is inherently cumulative and rigor is progressively greater (Garvin, 1973, p. 2).

2. Use for individual evaluation pertaining to competencies possessed by individual before instruction can be provided (Glaser and Nitko, 1971, p. 655).

3. Provides information for evaluating effectiveness of instruction based on instructional objectives (Hambleton and Gorth, 1971, p. 6).

IX. Limitations

A. Norm-referenced test

1. Limited to paper-and-pencil test (Cartier, 1968, p. 29).

B. Criterion-referenced test

1. Applicable to practical application such as putting a motor together, as well as paper-and-pencil test (Cartier, 1968, p. 29).

2. If intended to be comprehensive for a discipline and to cover material taught over several years, then the number of objectives needed to be represented on test becomes very large, and test would become excessively long (Roudabush and Green, 1971, p. 4).

3. Problems in number of items needed to measure objectives and in determining criterion of mastery.

Utility of Grade Equivalent Scores

Ahmann and Glock (1971) have suggested that grade equivalent scores are far more comprehensible to teachers, administrators, and the general public than are stanine and percentile scores reported as test results. Continuing their comments, the authors state:

> Grade equivalents offer convenient units for plotting profiles of student achievement. Such profiles are graphic representations of a pupil's test scores and typically emphasize the areas of overachievement and underachievement. Again, the reference point most useful for interpreting the profile is the present grade level of the pupil [p. 266].

Obtaining grade equivalent scores

Grade equivalent scores are frequently used to report test results because reporting the number of correct answers on a test via raw scores has little meaning by itself.

Grade equivalent scores are used with standardized reading tests which compare the achievement of a particular student with the achievement of a norming population. The norming population should include large numbers of students representing a variety of socioeconomic backgrounds and geographical areas. This population is given a preliminary form of the test, and, upon completion of the test administration, their scores are computed and norms are then determined. When the test is published, the norms are used to convert the raw scores into grade equivalent scores. A grade equivalent score of 2.6 refers to the raw score earned by an "average" second-grader at the sixth month of the school year. Thus, the grade equivalent score has greater utility than a raw score on a norm-referenced test.

Grade equivalent scores for reading are obtained by many of the following methods: (1) formal--standardized achievement tests, standardized reading tests, and standardized diagnostic reading tests, both oral and silent reading; and (2) informal--graded word lists, informal tests, and informal reading inventories.

Bond and Tinker (1967) describe how to make an informal reading inventory and the importance of grade equivalent indicators in so doing:

> Informal procedures can be accomplished through the use of a carefully graded series of basic readers. The series should be one which the child has not used before. Selections of 100 to 150 words are chosen from each successive book in the series. For any grade level, e.g.,

> grade 3.0, select material at about twenty pages from the beginning of the first book at that grade. Similarly, for halfway through a grade (grade 3.5, etc.) select material at about twenty pages from the beginning of the first book of that grade. A few questions involving both some ideas and some facts are constructed on each selection . . . . If the material in the book he starts with is not handled easily, he is moved back to a still easier level. The child then reads the successively more difficult selections until his reading levels are determined [p. 198].

As students are completing the above types of tests and inventories, kinds and number of errors are noted by the teacher or diagnostician, as well as their reading levels reported in approximate grade equivalent scores.

Using grade equivalent scores in reading

Grade equivalent scores can provide essential information for instructional decisions relating to (1) persons (both individuals and groups) and (2) materials to be used. For the purpose of providing an effective instructional program in reading, information is needed concerning the student's reading levels in order to supply the appropriate levels of materials required in the instructional process.

In terms of reading ability, three kinds of reading levels are usually reported for a student as follows: (1) independent reading level, (2) instructional reading level, and (3) frustration reading level. Bond and Tinker (1967) describe the reading levels as follows:

> 1. The child's independent reading level is ascertained from the book in which he can read with no more than one error in word recognition (pronunciation) in each 100 words and has a comprehension score of at least 90 percent.
>
> 2. The instructional reading level is determined from the level of the book in which the child can read with no more than one word-recognition error in each 20 words and has a comprehension score of at least 75 percent.
>
> 3. The frustration reading level is marked by the book in which the child "bogs down" when he tries to read [pp. 198-199].

Generally, the above reading levels are reported in grade equivalent scores which are useful in providing the appropriate levels of material for instruction and independent reading. In planning instruction, it is important to realize that achievement test scores for most individuals may be nearer their frustration level of reading than their instructional level, so adjustments in instructional materials may need to be made (Durkin, 1971, p. 410).

Reading levels are also reported in terms of a person's ability to read silently and/or orally, as the skills may differ considerably for an individual. Therefore, grade equivalent scores are frequently reported for silent reading and/or oral reading, depending upon the amount and nature of tests administered. Again, if this information is known, appropriate levels of materials can be supplied.

It is important to know what skills the student needs to master, but it is equally important to provide instruction for the student using the proper level of materials. Bond and Tinker (1967), in referring to remedial reading procedures, state that "selection of materials at the appropriate level of difficulty for a specific case is probably one of the most important decisions the diagnostician makes [p. 179]."

Grade equivalent scores are reported in determining an approximate reading expectancy level as developed by Harris (1970). The reading expectancy grade level is found by subtracting 5.2 from the individual's reading expectancy age. Therefore, if a student's reading expectancy age is 10.5, subtracting 5.2 from that figure produces a reading expectancy level of 5.3, which is the grade level at which the student is expected to read.

Grade equivalent levels are also utilized in readability formulas which indicate the approximate reading level in terms of grade level for individual books. Again, additional information is helpful in the selection of books to be used in instruction or to be read independently. The public schools of Kalamazoo, Michigan have utilized such a formula for the selection of textbooks in most subjects and have found that many textbooks supposedly written for a particular grade level of reading are totally unsuitable in terms of reading difficulty level for a particular grade.

For some students to function with any degree of success on a test, knowledge of the reading levels of the students and the reading level of the test must be known. As was noted in Chapter I, when a student misses an objective such as "indicate author's purpose" (which was an objective in the 1973-74 Michigan Educational Assessment Program given to all fourth-grade students in Michigan--whether they could read it or not), it is unclear whether the student has missed the item because the student could not "identify the author's purpose," or the selection was too difficult in terms of reading level, or whether other distracting factors were involved. Thus, knowledge of reading levels for both the student and the test is important information.

In summary, grade equivalent scores can offer essential information for decisions which involve classroom organization, grouping, difficulty level of instructional and testing materials, and an indication of general strengths and weaknesses of both individuals and classes. Thus, the importance and contribution of grade level scores in instructional planning decisions pertaining to persons and materials can be noted.

This chapter focused on a review of the literature as related to the history and development of criterion-referenced tests, characteristics of criterion-referenced tests, characteristics of norm-referenced tests, distinctions between norm-referenced tests and criterion-referenced tests, and the utility of grade equivalent scores in reading instruction.

Chapter III presents a brief review of the problem, description of the population and sample, instrumentation, procedures, and data treatment employed in the study.

CHAPTER III

DESIGN AND METHODOLOGY

The purpose of this chapter is to make explicit the design of the study and the procedure used to implement it. The following are explained: (1) review of the problem, (2) population and sample, (3) instrumentation, (4) procedures, and (5) data treatment.

Review of the Problem

The purpose of the present study was to determine what information can be provided by criterion-referenced tests to aid the educational leader in making instructional decisions. Specifically, the investigation determined whether or not information provided by the results of a criterion-referenced test can predict relative performance, i.e., approximate grade equivalent scores, as indicated on a norm-referenced test.

Population and Sample

The population for this investigation consisted of all students in the fourth and seventh grades in the Kalamazoo Public Schools, Kalamazoo, Michigan. The fourth-grade population was composed of 1,085 students from 42 classrooms in the 11 upper elementary schools in Kalamazoo.

The seventh-grade population was comprised of 1,168 students in the 5 junior high schools in the school system.

To be included in the sample, the students had to have completed both the criterion-referenced test and the norm-referenced test when administered. Thus, the fourth-grade sample consisted of 969 students, while the seventh-grade sample consisted of 949 students. For purposes of analyzing the data and making comparisons, the samples were divided into subgroups, as indicated in Table 1.

Table 1

Distribution of Students in Sample

Student Subgroup	Fourth Grade		Seventh Grade	
	No.	%	No.	%
Black	227	23.4	204	21.5
White	742	76.6	745	78.5
Male	491	50.7	492	51.8
Female	478	49.3	457	48.2
Black male	115	11.9	99	10.4
Black female	112	11.5	105	11.1
White male	376	38.8	393	41.4
White female	366	37.8	352	37.1
Total students	969	100.0	949	100.0

Instrumentation

Michigan Educational Assessment Program

The 1973-74 Michigan Educational Assessment Program

(MEAP) consisted of an objective-referenced reading test and mathematics test, at both the fourth- and seventh-grade levels. Both the fourth-grade and seventh-grade reading tests contained 5 test items for each of the 23 minimal objectives tested. The fourth-grade reading test contained some items which were read by the test administrator to the students.

The fourth-grade mathematics test contained 5 multiple-choice test items for each of the 35 minimal performance objectives tested, while the seventh-grade mathematics test contained 5 multiple-choice test items for each of the 45 objectives represented. Each test included several items to be read by the administrator.

Development of the 1973-74 MEAP by the Michigan Department of Education (1973) is explained in _School and District Reports: Explanatory Materials_, as follows:

> The minimal performance objectives were selected from those developed by educators and reviewed by commissions made up of teachers, administrators, curriculum specialists, and lay citizens cooperating with the Department of Education. Having been formally adopted by the State Board of Education, these performance objectives represent a set of minimal expectancies applicable to all beginning fourth and seventh grade students in Michigan. In 1972, a project was begun to write and validate test items to measure the minimal mathematics and reading objectives. Under contract to the Michigan Department of Education, five school districts (. . . in cooperation with the Michigan Council of Teachers of Mathematics) wrote the items. The test items were edited by staff members of the Department and of California Test Bureau/McGraw-Hill, who served as technical support contractor for the project.

> The items were tried out in the districts under contract and in the Detroit Public Schools. Final item revisions were based on teachers' comments and reviews by subject-matter specialists in the light of item tryout data. The final instruments were produced by the technical support contractor following specifications approved by the Department [pp. 2-3].

Metropolitan Achievement Test

The levels of the Metropolitan Achievement Test (MAT) administered and utilized in this study were the Elementary Level for fourth-grade students and the Advanced Level for seventh-grade students. The Metropolitan Achievement Test consists of 9 various subtests at the Elementary Level and 11 various subtests at the Advanced Level. However, for purposes of this study, the following 7 subtests in reading and mathematics were used at both the fourth- and seventh-grade levels: (1) word knowledge, (2) reading, (3) total reading, (4) mathematics computation, (5) mathematics concepts, (6) mathematics problem solving, and (7) total mathematics. Results of the above subtests were reported in grade equivalent scores for each student.

Development of the MAT has been described by Durost, Bixler, Wrightstone, Prescott, and Balow (1971), in _Metropolitan Achievement Test Teacher's Handbook_, as follows:

> _Metropolitan Achievement Tests_ are designed to evaluate what is being taught in today's schools. Therefore, the development of content for the tests depended on extensive analysis of current curricular materials. At the

> beginning of the test development effort, lists were made of leading textbook series, syllabuses, state guidelines, and other curricular sources. The test authors and authorial assistants next analyzed and summarized these materials. Based on these comprehensive summaries, test "blueprints" were prepared. The test blueprints indicated the proportion of test items on various topics needed to give balanced coverage to the curriculum . . . .
>
> After test blueprints were developed, the actual item writing took place. Items were written to cover each subtopic in the blueprints. A sufficient number of items was written for each subtopic so that, after classroom tryout, any items which were not functioning satisfactorily could be eliminated without adversely affecting the balance of test content. Following item writing, the items were edited by the publisher and reviewed by independent authorities. Appropriateness of content and format, clarity of wording, and other such factors were examined and, where possible, improved upon [pp. 15-16].

Validity for the MAT is somewhat tenuous, depending upon the curriculum of the schools using the tests. Durost et al. (1971) make the following comments concerning the validity of the tests:

> The validity of an achievement test is defined primarily in terms of _content validity._ A test has content validity if the test items adequately cover the curricular areas that the test is supposed to evaluate. Since each school has its own curriculum, the content validity of _Metropolitan Achievement Tests_ must be evaluated by each school. It cannot be claimed that the tests are universally valid. To assist schools in judging the content validity of the tests, the authors and publisher have prepared content outlines for the tests and described the procedures used in developing the test content [p. 16].

Reliability estimates for the 1970 edition of the Metropolitan Achievement Tests were determined by both split-half (odd-even) estimates corrected by the Spearman-Brown formula and Saupe's estimate of Kuder-Richardson Formula 20. Standard errors of measurement are given for grade equivalents and were determined using split-half coefficients. Both the reliability coefficients and standard errors of measurement for the Fall are based on data from all students who took Form G in the Fall standardization program and are included in Table 2.

MEAP and MAT

Since the subject areas of reading and mathematics were included in both MEAP (criterion-referenced test) and MAT (norm-referenced test), the information provided by the same subject areas on the two kinds of tests could be compared. Figure 1 indicates comparable subtests of MEAP and MAT, thus making comparisons possible.

MAT	MEAP
Reading subtests Word knowledge Reading Total reading	Reading (23 objectives)
Mathematics subtests Mathematics computation Mathematics concepts Mathematics prob. solving Total mathematics	Mathematics grade) for fourth (45 objectives for (45 objectives for seventh grade)
Scores reported in grade equivalent scores.	Scores reported by Y (objective mastered) and N (not mastered). Total number of objectives mastered for each individual on each subtest recorded for this study.

Fig. 1.--Plan for comparing the results of each MAT subtest with comparable MEAP subtest results.

Table 2

Reliability Data and Standard Errors of Measurement of Metropolitan Achievement Tests by Level and Grade for Fall Standardization

Subtest	Split-half Coefficient Corrected by Spearman-Brown Formula		Saupe's Estimate of Kuder-Richardson Formula 20		Standard Error of Measurement in Terms of Grade Equivalents	
	Elementary (4.1)	Advanced (7.1)	Elementary (4.1)	Advanced (7.1)	Elementary (4.1)	Advanced (7.1)
			Reading			
Word knowledge	.95	.92	.94	.91	.30	.60
Reading	.93	.92	.92	.91	.40	.60
Total reading	.97	.92	.96	.95	.30	.40
			Mathematics			
Computation	.91	.90	.88	.88	.30	.50
Concepts	.91	.87	.90	.87	.40	.70
Problem solving	.93	.90	.91	.89	.40	.60
Total mathematics	.97	.96	.96	.95	.20	.40

Procedures

The data used in this study were collected in September and October of 1973, in all fourth and seventh grades in the Kalamazoo Public Schools, Kalamazoo, Michigan. The Metropolitan Achievement Test, given to every fourth- and seventh-grade student, was administered by classroom teachers in both grades during the third week of September. The decision to administer the tests was made by the local school district in an effort to evaluate student achievement.

The Michigan Educational Assessment Program was also administered by classroom teachers to all students in both fourth and seventh grades during the first week of October. The decision to administer this test was made on the state level, in an effort to assess the academic achievement in reading and mathematics of fourth- and seventh-grade students in the State of Michigan.

The classroom teachers had administered the Metropolitan Achievement Test twice each year during the past few years; therefore, the administration procedures were familiar to them. However, workshops concerning MAT testing procedures were held in each school, by the school's testing coordinator, for the teachers. The Michigan Educational Assessment Program was new to the classroom teachers; therefore, extensive workshops were given for the teachers in each school by the testing coordinator in the school, explaining the nature of MEAP and the testing administration procedures for the test.

The student response sheets from both tests were sent to the appropriate company or agency for machine scoring, and student and class summaries of the test results were returned to the Department of Research and Development of the Kalamazoo Public Schools for distribution to the classroom teachers.

For this investigation, the results of the Metropolitan Achievement Test and the Michigan Educational Assessment Program were processed by first counting and recording, for each student, the number of reading objectives mastered and mathematics objectives mastered on the MEAP.

Next, for each student in the fourth-grade sample (N = 969) and for each seventh-grade student (N = 949), the following summary information was placed on an appropriate form, to be processed for computer usage:

1. Student code number
2. MAT reading subtest, reported in grade equivalent scores:
   a. Word knowledge
   b. Reading
   c. Total reading
3. MAT mathematics subtests, reported in grade equivalent scores:
   a. Computation
   b. Concepts
   c. Problem solving
   d. Total mathematics
4. Number of MEAP reading objectives mastered.

5. Number of MEAP mathematics objectives mastered.

The above information was coded for key punching on IBM cards, and then the completed IBM cards were proofread and all errors were corrected. Existing computer programs were adapted to meet requirements of the data treatment.

Data Treatment

The following statistical methods were used in treatment of the data. An explanation of each of the treatments follows:

Histograms and scatter diagrams.--To determine the nature of the distributions of the results from the MEAP Reading Test and Mathematics Test, histograms were made showing the total number of objectives passed by each student and the percentage of the sample achieving that number. Four histograms were made showing the distributions for both the MEAP reading and mathematics tests, for grades four and seven.

Scatter diagrams were used to plot the relationship between the total number of objectives passed on each MEAP test with the grade equivalent scores from the corresponding subtests on the MAT for each individual. This procedure was used for both the reading and mathematics grade equivalent scores from the MAT, for grades four and seven.

Descriptive data.--For each of the subgroups in both grades four and seven, the following were calculated:

mean, standard deviation, variance, median, mode, standard error of the mean, and the coefficient of variance.

Correlation coefficients.--Due to the nature of the results of the histograms and the scatter diagrams, three different correlation coefficients were used. The correlation coefficients were computed to show the relationship between the total number of objectives passed on a test from the MEAP and the grade equivalent scores from the corresponding tests from the MAT. This was computed for all subgroups on all subtests in both fourth and seventh grades. The coefficients of determination ($\underline{r}^2$ and $R_Q{}^2$) were also reported.

Two different types of correlation measures were employed:

1. Pearson ($\underline{r}$), which shows the degree of linear relationship.
2. Index of Correlation (R_Q) based on a quadratic model, which indicates the degree of relationship explained by a curved line (Wert, Neidt, and Ahmann, 1954, ch. 15).

Comparison of coefficients of determination.--Procedures to compare the coefficients of determination were used to estimate the percentage of variance increase due to the addition of the quadratic term.

Regression equation.--Predicted grade equivalent scores were computed using the following formula:

$$Y = b_0 + b_1(X) = b_2(X^2)$$

Z Transformation of correlation coefficients.--The

Index of Correlation values from all subtests and most subgroups were transformed into Z values by the formula

$$Z = \frac{1}{2} \log_e \left[\frac{1 + r}{1 - r}\right]$$

The Z values were averaged and changed to the corresponding correlation coefficients, using an appropriate logarithm table constructed by Malloy, and using the above formula. With the resultant correlation coefficient values, a summary table of the results of the study was constructed, to be used as a basis for discussion of the findings of the study.

Services of the computer facilities of Western Michigan University and the Department of Research and Development of the Kalamazoo Public Schools were utilized.

In Chapter III, a concise problem statement was reviewed, and the population and sample used in the study were identified. Instrumentation and procedures utilized in the study were described, with a section on data analysis concluding the chapter.

Research findings are reported and discussed in Chapter IV.

CHAPTER IV

REPORT OF THE FINDINGS

In this chapter, the findings are reported as they relate to each of the following questions which were posed for investigation in Chapter I:

1. What is the relationship between norm-referenced tests and criterion-referenced tests with respect to predicting student performance scores or grade equivalents as indicated on a norm-referenced test?

2. What is the relationship between scores on norm-referenced tests and criterion-referenced tests for (a) fourth- and seventh-grade students, (b) black students and white students, and (c) male and female students?

The primary statistical models used to analyze the data were linear and nonlinear regression procedures which were appropriate as indicated by scatter diagrams.

Data presentation consists of reporting the coefficients of correlation ($\underline{r}$ and R_Q), comparisons of the coefficients of determination ($\underline{r}^2$ and $R_Q{}^2$), and the estimated probability (p) of the observed relationships being a chance occurrence.

Question One

What is the relationship between norm-referenced tests and criterion-referenced tests with respect to predicting student performance scores or grade equivalents as indicated on a norm-referenced test?

The degree of relationship between the 1973-74 Michigan Educational Assessment Program (MEAP) scores and the Metropolitan Achievement Test (MAT) scores was large and consistent. This relationship was replicated across subjects and tests, as shown in Table 3 and Appendix A. It is clear with a high degree of confidence that the relationships were highly significant, since the p values were less than .0001. Since a relatively high relationship does exist between the MEAP (criterion-referenced test) and the MAT (norm-referenced test), it is possible that a criterion-referenced test can predict student performance scores as grade equivalents.

Table 3

Summary Data for Relationship Between Scores on Metropolitan Achievement Test and Michigan Educational Assessment Program in Reading and Mathematics

Subtest	Fourth Grade		Seventh Grade	
	r	R_Q	*r*	R_Q
<u>Reading</u>				
Word knowledge	.68	.70	.70	.72
Reading	.70	.73	.76	.77
Total reading	.71	.74	.76	.77
<u>Mathematics</u>				
Computation	.57	.62	.72	.77
Concepts	.62	.71	.75	.80
Problem solving	.59	.68	.73	.79
Total mathematics	.64	.71	.79	.84

$p < .0001$

Scatter diagram indications

Scatter diagrams (Appendix B) were utilized to show the relationship of the total number of objectives passed for each individual on the MEAP with the same individual's corresponding grade equivalent score on the MAT. For instance, the total number of objectives passed by an individual on the MEAP Reading Test was correlated with the same individual's grade equivalent score received on the MAT Total Reading Subtest. The same procedure was used in showing the relationship in the area of mathematics for all students in the fourth- and seventh-grade samples.

The scatter diagrams indicate a relationship between the Michigan Educational Assessment Program and the Metropolitan Achievement Test. However, the relationships were not always linear. The scatter diagrams indicated a greater amount of linearity in the reading tests at both the fourth- and seventh-grade levels than on either the fourth-grade or seventh-grade mathematics tests.

Comparison of coefficients of determination

The following procedures were utilized to determine the percentage of increase in the strength of the relationship between the scores of the MEAP and MAT attributed to the addition of the squared variable represented in the quadratic model. Both the linear (*r*) and the quadratic

(R_Q) correlation coefficients were computed. After the linear coefficient of determination ($\underline{r}^2$) and the quadratic ($R_Q{}^2$) were computed, they were then subtracted $[(R_Q{}^2) - \underline{r}^2]$ to find the percentage of increase in the variance which was due to the quadratic term alone.

A test used in stepwise regression, to determine if an additional independent variable should be added to a general linear model, was adopted and used. This test determined whether the percentage of increase due to the quadratic alone was significant. The formula is

$$F = (N - 3)\frac{(R_Q)^2 - (r)^2}{1 - (R_Q)^2}$$

The degrees of freedom used with this formula are 1 and N - 3 where N is the sample size. After the $\underline{F}$ values were computed and compared with an $\underline{F}$ distribution table, it was concluded that all values in the column "Increase Due to Quadratic Alone" in Tables 4 and 5 were significant at the .01 level.

Although the differences in the linear and quadratic correlation coefficients were small, the significant differences noted on comparing the coefficients of determination for the linear and quadratic models were probably due to the large sample sizes ($N = 969$, fourth grade; $N = 949$, seventh grade). Based on the above information, it can be concluded that the use of the quadratic model (R_Q) could be considered to be more appropriate than the linear model.

Table 4

Increase of Variance Due to Quadratic Model Computed to Linear Model by Use of Coefficients of Determination (Fourth Grade)

Subtest	Linear	Quadratic	Coefficient of Determination of Linear	Coefficient of Determination of Quadratic	Increase Due to Quadratic Alone
	$\underline{r}$	R_Q	$\underline{r}^2$	$R_Q{}^2$	$(R_Q)^2 - \underline{r}^2$
			Reading		
Word knowledge	.683	.702	.4665	.4928	.0263*
Reading	.696	.730	.4844	.5329	.0485*
Total reading	.709	.736	.5027	.5417	.0390*
			Mathematics		
Computation	.572	.624	.3272	.3894	.0622*
Concepts	.619	.705	.3832	.4970	.1138*
Problem solving	.592	.683	.3505	.4665	.1160*
Total mathematics	.635	.710	.4032	.5041	.1009*

*$p < .01$

Table 5

Increase of Variance Due to Quadratic Model Computed to Linear Model by Use of Coefficients of Determination (Seventh Grade)

Subtest	Linear	Quadratic	Coefficient of Determination of Linear	Coefficient of Determination of Quadratic	Increase Due to Quadratic Alone
	r	R_Q	r^2	${R_Q}^2$	$(R_Q)^2 - r^2$
			Reading		
Word knowledge	.700	.724	.4900	.5242	.0342*
Reading	.756	.772	.5715	.5960	.0245*
Total reading	.755	.776	.5700	.6022	.0322*
			Mathematics		
Computation	.717	.771	.5141	.5944	.0803*
Concepts	.753	.798	.5670	.6368	.0698*
Problem solving	.730	.792	.5329	.6273	.0944*
Total mathematics	.786	.839	.6178	.7039	.0861*

*$p < .01$

Also, it can be noted in Tables 4 and 5 that the estimated percentage of variance increase due to the addition of the quadratic term is larger for arithmetic (6 to 11.6 percent) than for reading (2.5 to 4.9 percent). As in the case of arithmetic, a 10-percent increase in variance is large or substantial; therefore, the relationship is more nonlinear between the variables for arithmetic than for reading.

Question Two

What is the relationship between scores on norm-referenced tests and criterion-referenced tests for (a) fourth- and seventh-grade students, (b) black students and white students, and (c) male and female students?

Fourth- and seventh-grade students

As can be noted in Table 6, the corresponding correlation coefficients for all subgroups in both reading and mathematics were higher for the seventh-grade sample than for the fourth-grade sample. While the differences between the reading correlation coefficients across grade levels were quite small (.01 to .08) and consistent, the mathematics correlation coefficients exhibited greater variability (.04 to .18) in the differences across subgroups and grade levels. For instance, for most subgroups the correlation coefficients in mathematics in grade seven were

Table 6

Quadratic Correlation Coefficients for All Subgroups in Reading and Mathematics for Grades Four and Seven

Subgroup	Grade Four		Grade Seven	
	Reading R_Q	Math R_Q	Reading R_Q	Math R_Q
All student	.72	.68	.76	.80
Black student	.69	.63	.70	.63
White student	.70	.63	.72	.79
Male student	.73	.67	.75	.81
Female student	.72	.70	.77	.81

$p < .0001$

from .11 to .16 higher than were the fourth-grade mathematics correlation coefficients, except for the black student subgroup. In both grades four and seven, the black student subgroup had a lower correlation coefficient of .63 in mathematics, as compared to an identical mathematics correlation coefficient for the white students in grade four, but a high correlation coefficient of .79 was reported for the white students in mathematics in grade seven. Thus, a higher seventh-grade mathematics correlation coefficient for the black student subgroup was not indicated as it was with the remaining subgroups.

Within the fourth-grade sample, the reading correlation coefficients were higher for all subgroups of the sample than were the mathematics correlation coefficients. In the area of reading, the highest correlation coefficient was reported for the male student subgroup, while the lowest correlation coefficient was reported for the black student subgroup. It must be noted, however, that the correlation coefficients were very consistent, ranging from .69 to .73, with only a small amount of variability noted.

Within the fourth-grade sample, the mathematics correlation coefficients (.63 to .70) were generally lower than were the reading correlation coefficients, with slightly greater variability noted. In the area of mathematics, the female students reported the highest correlation coefficients of .70, with both the black students and white students reporting the same correlation coefficient (.63), which was the lowest reported.

Within the seventh-grade sample, the reading correlation coefficients ranged from .70 to .77, with the lowest correlation coefficient reported for the black student subgroup and the highest correlation coefficient in reading reported for the female student subgroup.

Within the seventh-grade sample, the mathematics correlation coefficients (.63 to .81) had greater variability than any of the other fourth-grade and seventh-grade reading and mathematics correlation coefficients reported. The high correlation coefficient of .81 in mathematics for the male student and female student subgroups was the highest correlation coefficient reported across grades and tests. Again, the lowest correlation coefficient of .63 was reported for the black student subgroup.

Black students and white students

In an analysis of correlation coefficients according to racial subgroups, the white students reported higher correlation coefficients than the black students across tests and grades. At the fourth-grade level, both black students and white students reported nearly the same correlation coefficients (.69 and .70) in reading, and the same correlation coefficient (.63) in mathematics. In the seventh-grade sample, however, a small difference in correlation coefficients (.70 and .72) was reported in reading, while a wide discrepancy in correlation coefficients (.63 and .79) was reported between races in the area of mathematics. For example, the white student subgroup correlation coefficient indicating the relationship between the MAT and MEAP mathematics scores was .79, which is relatively high, while the correlation coefficient for the black students between the same two tests was only .63. Thus, the seventh-grade mathematics correlation coefficients between black students and white students constituted the greatest discrepancy reported across tests and subgroups.

Male and female students

The highest correlation coefficient reported between the MAT and MEAP scores was for the female student subgroup and the all-student group. The female students reported higher correlation coefficients by a difference of .01 or .02 on all tests for both grades, except for the fourth-grade reading correlation coefficient in which the female student subgroup was next highest to the male student subgroup. The fourth-grade reading correlation coefficients for both the all-student subgroup and the female student subgroup were identical at .72.

In an analysis of subgroup correlation coefficients in both reading and mathematics for both grades, sex (male and female) had higher correlation coefficients between the MAT and MEAP scores than did race (white students and black students), with higher correlation coefficients reported for females than for males. Thus, among all the subgroups the correlation coefficients reported for the females were the highest across grades and tests, with the male student subgroup reporting the next highest correlation coefficients.

In summary, the higher correlation coefficients between grades were reported for the seventh-grade sample. The highest correlation coefficients across grades and tests were reported for the female student subgroup, while the lowest correlation coefficients were reported for the black student subgroup.

In the areas of reading and mathematics for the fourth grade, higher correlation coefficients were reported in reading than in mathematics. On the other hand, in the areas of reading and mathematics for the seventh grade, the higher correlation coefficients were reported in mathematics rather than in reading.

Predicted Grade Equivalent Scores

Since the results of the findings indicate that approximate grade equivalent scores can be predicted, Table 7 is presented to illustrate the possibilities for use. Table 7 gives the predicted grade equivalent scores from the number of objectives mastered on the fourth-grade MEAP (criterion-referenced test) in the area of reading. The following formulas (Ezekial and Fox, 1959, chs. 14-15) could be used to compute the predicted grade equivalent scores:

Linear Model $Y = b_0 + b_1(X)$

Example $(X = 1)$ $Y = 2.9 + .13(1) = 3.03$

Quadratic Model $Y = b_0 + b_1(X) + b_2(X^2)$

Example $(X = 1)$ $Y = 2.5 + .03(1) + .007(1) = 2.537$

In the above formulas, the symbols represent the following:

Y = predicted grade equivalent score

b_0 = estimated constant term

b_1 = estimated linear regression term

b_2 = estimated quadratic regression term

X = total number of objectives mastered on the criterion-referenced test

Table 7

Predicted Grade Equivalent Scores from Number of Objectives Mastered on Michigan Educational Assessment Program Reading Test (Grade Four)

Number of Objectives Mastered	Grade Equivalent Score[a]	Number of Objectives Mastered	Grade Equivalent Score[a]
1	2.5	13	4.1
2	2.6	14	4.3
3	2.7	15	4.5
4	2.7	16	4.8
5	2.8	17	5.0
6	2.9	18	5.3
7	3.0	19	5.6
8	3.2	20	5.9
9	3.3	21	6.2
10	3.5	22	6.5
11	3.7	23	6.9
12	3.9		

[a] Based on quadratic model $Y = b_0 + b_1(X) + b_2(X^2)$

Standard error of estimate = 1.13

This chapter presented the results of the study by reporting the findings for question one, which included discussion of the scatter diagram indications and the correlation coefficients. In the discussion relating to question two, the findings for the following subgroups were presented: fourth grade and seventh grade, black students and white students, male and female students.

The chapter was concluded by charting examples of predicted grade equivalent scores related to the study.

Chapter V contains a summary of the study, conclusions and discussion pertaining to the findings, and recommendations for possible application and future research.

CHAPTER V

SUMMARY, CONCLUSIONS, AND IMPLICATIONS

Summary

This study can be focused on discovering what information can be provided by criterion-referenced tests to aid the educational leader in making instructional decisions. Specifically, it was the intent of the investigator to determine whether or not information provided by the results of a criterion-referenced test can predict relative performance, i.e., approximate grade equivalent scores, as indicated on a norm-referenced test.

In order to complete the specific objectives of the study, three major questions were investigated:

1. What is the relationship between norm-referenced tests and criterion-referenced tests with respect to predicting student performance scores or grade equivalents as indicated on a norm-referenced test?

2. What is the relationship between scores on norm-referenced tests and criterion-referenced tests for (a) fourth- and seventh-grade students, (b) black students and white students, and (c) male and female students?

3. What information can the criterion-referenced test provide educational decision makers in decisions pertaining to placement, diagnosis, assessment, prediction, and evaluation?

The sample consisted of 969 students from the 1,085 fourth-grade population representing 42 classrooms in the 11 upper elementary schools and 949 students from the 1,168 seventh-grade population representing the 5 junior high schools in the public school system of Kalamazoo, Michigan. The sample was limited to those students who were present for the administration of both tests used in the study.

Data were collected by classroom teachers administering the Metropolitan Achievement Test (norm-referenced test) and the 1973-74 Michigan Educational Assessment Program (criterion-referenced test) to the same students in late September and early October of 1973. The data were summarized and keypunched on IBM cards, for processing on the computer system at Western Michigan University.

Data analysis consisted of the following:

1. Constructing histograms and scatter diagrams to observe the nature of the distributions of the data.

2. Calculating descriptive information: mean, standard deviation, variance, median, mode, standard error of the mean, and the coefficient of the variance.

3. Calculating correlation coefficients: Pearson and Index of Correlation.

4. Transforming correlation coefficients to Z scores, averaging, and then converting to correlation coefficients for final summary table of data.

5. Predicting approximate grade equivalent scores by using the appropriate regression equation.

Conclusions

Two of the specific questions investigated in this study were presented and analyzed in Chapter IV. Conclusions and discussion of the results of the analysis are presented for each of the questions posed for investigation.

Question one

What is the relationship between norm-referenced tests and criterion-referenced tests with respect to predicting student performance scores or grade equivalents as indicated on a norm-referenced test?

The degree of relationship between the MEAP scores and the MAT scores was relatively high and consistent. This relationship was replicated across subjects and tests. It is clear with a high degree of confidence that the relationships were highly significant, since the p values were less than .0001. Therefore, since a relatively high relationship does exist between the MEAP (criterion-referenced test) and the MAT (norm-referenced test), it is possible that a criterion-referenced test can predict student performance as grade equivalents. For an example of predicted grade equivalent scores, refer to Table 7 (p. 65) in Chapter IV.

Question two

What is the relationship between scores on norm-referenced tests and criterion-referenced tests for (a) fourth- and seventh-grade students, (b) black students and white students, and (c) male and female students?

Fourth- and seventh-grade students.--The predictability of approximate grade equivalent scores is not the same for both fourth and seventh grades, as the seventh-grade correlation coefficients were higher for both tests (reading and mathematics) and for all subgroups. Therefore, the predictability of grade equivalent scores is higher, to some extent, for seventh grade. The mathematics test at the seventh-grade level reported generally higher correlation coefficients than did the reading test at the same grade level.

Black students and white students.--The greatest discrepancies and variability in correlation coefficients were in the racial subgroups. The black student subgroup reported the lowest correlation coefficients in both reading and mathematics at both the fourth- and seventh-grade levels, with the largest discrepancy between races reported in the mathematics test at the seventh-grade level. A correlation coefficient of .79 was reported for the white student subgroup on the mathematics test at the seventh-grade level, while .63 was reported for the black student subgroup on the same test in seventh grade.

The lower correlation coefficients for the black student subgroup might have been influenced by their lower achievement levels, as indicated in Table 8.

Table 8

Mean Grade Equivalent Scores on Metropolitan Achievement Test Given in September, 1973

Subtest	Fourth Grade		Seventh Grade	
	Black Subgroup (N = 227)	White Subgroup (N = 742)	Black Subgroup (N = 204)	White Subgroup (N = 745)
Total reading	2.7	4.1	5.0	7.4
Total mathematics	3.1	4.1	5.6	7.3

Since the MEAP (criterion-referenced test) was written to represent the minimal skills of the respective areas of reading and mathematics in both the fourth and seventh grades and the mean grade equivalent achievement of the black students on the MAT was nearly one or two years below the grade level of materials represented on the MEAP, perhaps the MEAP was inappropriate for the black student subgroup. If so, considerable guessing may have taken place, resulting in more erratic responses for the black student subgroup than for the white student subgroup, thus resulting in lower correlation coefficients.

Male and female students.--Little difference (.01 to .03) was noted between the correlation coefficients for the male and female students across tests and grades. In the area of reading at the fourth-grade level, the males reported a correlation coefficient of .73, which was .01

higher than the female correlation coefficient of .72; the females had a correlation coefficient of .77, .02 higher than the males (.75) in reading at the seventh-grade level. In mathematics, the females reported a correlation coefficient of .70 at the fourth-grade level, which was .03 higher than the male correlation coefficient of .67. However, both male and female student subgroups reported a correlation coefficient of .81 in mathematics at the seventh-grade level. In general, the correlation coefficients for the male and female students were very similar, with little difference reported. Therefore, the predictability of grade equivalent scores based on sex is approximately the same.

Implications

Question three

What information can the criterion-referenced test provide educational decision makers in decisions pertaining to placement, diagnosis, assessment, prediction, and evaluation?

According to Katz (1972), the above classifications are the "intermediate" purposes of testing, and this study attempted to discover what information a criterion-referenced test can provide to achieve these purposes. Based on the results of this study, it may be considered possible for a criterion-referenced test to (1) specify objectives mastered and/or not mastered and (2) indicate an approximate grade level score.

Placement refers to placing students in relation to one another in various groups (selection) and, also, to placing a student at an appropriate level in an instructional sequence of content in a subject. Both of these placement functions can be met by using a criterion-referenced test in educational decisions of placement if the criterion-referenced test can provide both an approximate grade level score and specific objectives mastered/not mastered. For example, in placing students in relation to one another in various groups (selection), the students could be placed in groups based upon their approximate grade level scores and then be grouped for detailed instruction according to their specific needs as indicated on the objectives of the criterion-referenced test. Therefore, the approximate grade equivalent score would form the basis for initial grouping, and then the objectives mastered/not mastered would provide the basis for grouping within the group for purposes of instruction relating to specific objectives.

Approximate grade equivalent scores and performance objectives mastered/not mastered can be very important information in matching materials and a sequence of materials to the student's instructional needs. Instructional materials, especially in reading, are carefully sequenced according to grade level difficulty, so suitable levels of materials can be utilized in the instructional process. Since many reading objectives are skills which are developed over many grade levels, it is essential to know the appropriate grade level of materials to use in teaching the desired objectives to a particular student.

Since both grade equivalent scores and performance objectives mastered/not mastered can be provided by criterion-referenced tests, the tests can be utilized for placement purposes of both persons and materials within a school, or within a classroom or small instructional groups, or within an instructional unit of study in a specific subject area.

Diagnosis involves analyzing in depth the strengths and weaknesses of particular students regarding their skills, knowledge, and style of learning. Since criterion-referenced tests can provide information pertaining to both approximate level of functioning and specific descriptive information on skills mastered/not mastered, the criterion-referenced test is especially applicable to diagnosis, as both kinds of information are essential in planning instruction for the student utilizing the appropriate level of materials and knowing which skills need to be worked on and which ones are the student's points of strength.

Criterion-referenced test information can also be used in diagnosing the needs of a whole class as well as the individual needs within a class. Since it is possible that a teacher may be instructing his students in materials which are more difficult than most students can handle effectively, approximate grade level information is important to the diagnosis of both class and individual needs.

Assessment involves measuring the effectiveness of a teaching method or treatment and can be utilized to indicate the amount of student growth and development. Both approximate grade level indicators and skill needs are important in this purpose of testing. On a criterion-referenced pretest, the information can give baseline data as well as giving information for the instruction of the students by providing the approximate grade level of materials needed and the skills needed to be mastered. On a posttest, both the comparisons of skills mastered and approximate grade equivalent growth could be measured.

Prediction utilizes measurement for the purpose of forecasting an individual's future performance on the basis of test results. Again, since criterion-referenced tests can provide both a "skill-needs" evaluation and an approximate grade level performance, the criterion-referenced test (which can also provide approximate grade level indicators) can provide more information for forecasting than either the norm-referenced test or criterion-referenced test alone has done in the past.

Evaluation, according to Katz (1972), involves the use of tests to compare one school with other comparable schools. If one wishes to compare schools based upon test results, perhaps only a broad or more general type of information is needed; therefore, this could be done as it has been traditionally, through the use of grade equivalent scores. If criterion-referenced tests are used and most all students are tested (as opposed to a random sample), then important information pertaining to the skill needs of individuals, classes, and schools can be analyzed and can form a basis for long-term as well as short-term instructional and curriculum planning at all levels.

In the past, choices had to be made between criterion-referenced and/or norm-referenced tests to supply adequate information for the above purposes. It is now possible to use a criterion-referenced test alone, if it provides information pertaining to specific objectives mastered/not mastered and grade equivalent scores for each of the above five purposes of testing.

REFERENCES

Ahmann, J. S., and Glock, M. D. *Evaluating pupil growth.* (4th ed.) Boston: Allyn and Bacon, 1971.

Beall, J. G., Jr. March 22, 1973. *Congressional Record,* 119 (45), S 5373.

Bond, G. L., and Tinker, M. A. *Reading difficulties: Their diagnosis and correction.* (2nd ed.) New York: Appleton-Century-Crofts, 1967.

Cartier, F. A. Criterion-referenced testing of language skills. *Washington TESOL Quarterly,* 1968, 2 (1).

Committee on Education and Labor, House of Representatives, U.S. Congress. Memo to Education Writers, Education Associations, Chief State School Officers, and School Board Members, August 15, 1973.

Cronbach, L. J. *Essentials of psychological testing.* (3rd ed.) New York: Harper & Row, 1970.

Davis, F. B. Criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York, February 1971.

Durkin, D. *Teaching them to read.* Boston: Allyn and Bacon, 1971.

Durost, W. N., Bixler, H. H., Wrightstone, J. W., Prescott, G. A., and Balow, I. H. *Metropolitan achievement test teacher's handbook.* New York: Harcourt Brace Jovanovich, 1971.

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

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

Ezekiel, M., and Fox, K. A. *Methods of correlation and regression analysis.* New York: John Wiley & Sons, 1966.

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

Frymier, J. R. Foreword. In A. W. Combs, Educational accountability: Beyond behavioral objectives. Washington, D.C.: Association for Supervision and Curriculum Development, 1972.

Garvin, A. D. The applicability of criterion-referenced measurement by content area and level. Paper presented at a joint session of the American Educational Research Association and National Council on Measurement in Education annual meetings, Minneapolis, March 1970.

Glaser, R. Instructional technology and the measurement of learning outcomes. In W. J. Popham (Ed.), Criterion-referenced measurement. Englewood Cliffs, N.J.: Educational Technology Publications, 1973.

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

Gronlund, N. C. Measurement and evaluation in teaching. New York: Macmillan, 1971.

Gronlund, N. C. Preparing criterion-referenced tests for classroom instruction. New York: Macmillan, 1973.

Hambleton, R. K., and Gorth, W. P. Criterion-referenced testing: Issues and applications. A version of this paper was presented at the annual meeting of the Northwestern Educational Research Association, Liberty, New York, 1970.

Harris, A. J. How to increase reading ability. (5th ed.) New York: David McKay, 1970.

Katz, A. Selecting an achievement test: Principles and procedures. In V. H. Noll, D. P. Scannell, and R. P. Noll (Eds.), Introductory readings in educational measurement. Boston: Houghton Mifflin, 1972.

Kirby, T. J. Practice in the case of school children. New York: Columbia University, Teachers College, 1913.

Klein, S. Evaluating tests in terms of the information they provide. University of California at Los Angeles, Center for the Study of Evaluation. Evaluation Comment, 1970, 2 (2).

Michigan Department of Education, Research, Evaluation, and Assessment Services. School and district reports: Explanatory materials. (Third report of the 1973-74 Michigan Educational Assessment Program) Lansing, Mich.: Author, 1973.

Nitko, A. J. Criterion-referenced testing in the context of instruction. Paper presented at the Educational Records Bureau-National Council on Measurement in Education Symposium, New York, October 1970.

Nunnally, J. C. Educational measurement and evaluation. (2nd ed.) New York: McGraw-Hill, 1972.

Popham, W. J., and Husek, T. R. Implications of criterion-referenced measures. In W. J. Popham (Ed.), Criterion-referenced measurement. Englewood Cliffs, N.J.: Educational Technology Publications, 1973.

Roudabush, G. E. Item selection for criterion-referenced tests. Paper presented at the meeting of the American Educational Research Association, New Orleans, February 1973.

Roudabush, G. E., and Green, D. R. Some reliability problems in a criterion-referenced test. Paper presented at the annual meeting of the American Educational Research Association, New York, February 1971.

Thorndike, E. L. The original nature of man. Vol. 1. Educational psychology. New York: Columbia University, Teachers College, 1913.

Wert, J. E., Neidt, C. O., and Ahmann, J. S. Statistical methods in educational and psychological research. New York: Appleton-Century-Crofts, 1954.

APPENDIX A

Table A

Relationship Between Michigan Educational Assessment Program Scores and Metropolitan Achievement Test Scores for Fourth- and Seventh-Grade Students

Subtest	Fourth Grade				Seventh Grade			
	Linear		Quadratic		Linear		Quadratic	
	r	r^2	R_Q	R_Q^2	r	r^2	R_Q	R_Q^2
				Reading				
Word knowledge	.68[a]	.47[b]	.70[c]	.49[d]	.70[a]	.49[b]	.72[c]	.52[d]
Reading	.70	.49	.73	.53	.76	.57	.77	.60
Total reading	.71	.50	.74	.54	.76	.57	.78	.60
				Mathematics				
Computation	.57	.33	.62	.39	.72	.51	.77	.59
Concepts	.62	.38	.71	.50	.75	.57	.80	.64
Problem solving	.59	.35	.68	.47	.73	.53	.79	.63
Total mathematics	.64	.40	.71	.50	.79	.62	.84	.70

[a] Linear r in this column = Pearson correlation coefficient.

[b] Linear r^2 in this column = coefficient of determination.

[c] Quadratic R_Q in this column = Index of Correlation.

[d] Quadratic R_Q^2 in this column = coefficient of determination.

$p < .0001$

Table B

Relationship Between Michigan Educational Assessment Program Scores and Metropolitan Achievement Test Scores for Fourth-Grade Black Students and White Students

Subtest	Black (N = 227)				White (N = 742)			
	Linear		Quadratic		Linear		Quadratic	
	$\underline{r}$	$\underline{r}^2$	R_Q	R_Q^2	$\underline{r}$	$\underline{r}^2$	R_Q	R_Q^2
Reading								
Word knowledge	.67[a]	.45[b]	.68[c]	.47[d]	.65[a]	.42[b]	.67[c]	.45[d]
Reading	.67	.45	.69	.48	.67	.45	.71	.50
Total reading	.69	.47	.70	.50	.68	.46	.71	.51
Mathematics								
Computation	.55	.30	.58	.33	.53	.28	.59	.35
Concepts	.60	.35	.66	.44	.58	.34	.67	.45
Problem solving	.53	.28	.61	.37	.56	.31	.65	.43
Total mathematics	.63	.40	.68	.47	.59	.35	.67	.45

[a] Linear $\underline{r}$ in this column = Pearson correlation coefficient.

[b] Linear $\underline{r}^2$ in this column = coefficient of determination.

[c] Quadratic R_Q in this column = Index of Correlation.

[d] Quadratic R_Q^2 in this column = coefficient of determination.

$p < .0001$

Table C

Relationship Between Michigan Educational Assessment Program Scores and Metropolitan Achievement Test Scores for Fourth-Grade Male and Female Students

Subtest	Male (N = 491)				Female (N = 478)			
	Linear		Quadratic		Linear		Quadratic	
	$\underline{r}$	$\underline{r}^2$	R_Q	R_Q^2	$\underline{r}$	$\underline{r}^2$	R_Q	R_Q^2
Reading								
Word knowledge	.68[a]	.47[b]	.70[c]	.48[d]	.70[a]	.46[b]	.71[c]	.51[d]
Reading	.72	.52	.74	.55	.67	.45	.72	.52
Total reading	.74	.54	.75	.56	.68	.46	.72	.52
Mathematics								
Computation	.59	.35	.62	.39	.55	.30	.63	.40
Concepts	.63	.40	.70	.49	.60	.36	.72	.51
Problem solving	.59	.35	.67	.45	.60	.35	.70	.49
Total mathematics	.64	.41	.70	.48	.62	.39	.73	.53

[a] Linear $\underline{r}$ in this column = Pearson correlation coefficient.

[b] Linear $\underline{r}^2$ in this column = coefficient of determination.

[c] Quadratic R_Q in this column = Index of Correlation.

[d] Quadratic R_Q^2 in this column = coefficient of determination.

$p < .0001$

Table D

Relationship Between Michigan Educational Assessment Program Scores and Metropolitan Achievement Test Scores for Fourth-Grade Male and Female Black Students

Subtest	Male (N = 115)				Female (N = 112)			
	Linear		Quadratic		Linear		Quadratic	
	r	r^2	R_Q	R_Q^2	r	r^2	R_Q	R_Q^2
Reading								
Word knowledge	.70[a]	.46[b]	.68[c]	.46[d]	.66[a]	.44[b]	.69[c]	.48[d]
Reading	.71	.51	.72	.53	.63	.40	.67	.44
Total reading	.73	.53	.73	.54	.65	.42	.68	.47
Mathematics								
Computation	.56	.32	.57	.33	.53	.29	.59	.34
Concepts	.59	.35	.65	.43	.60	.36	.69	.47
Problem solving	.52	.27	.58	.34	.55	.30	.63	.40
Total mathematics	.65	.42	.68	.46	.61	.38	.69	.47

[a] Linear r in this column = Pearson correlation coefficient.

[b] Linear r^2 in this column = coefficient of determination.

[c] Quadratic R_Q in this column = Index of Correlation.

[d] Quadratic R_Q^2 in this column = coefficient of determination.

$p < .0001$

Table E

Relationship Between Michigan Educational Assessment Program Scores and Metropolitan Achievement Test Scores for Fourth-Grade Male and Female White Students

Subtest	Male (N = 376)				Female (N = 366)			
	Linear		Quadratic		Linear		Quadratic	
	r	r^2	R_Q	R_Q^2	r	r^2	R_Q	R_Q^2
	Reading							
Word knowledge	.66[a]	.44[b]	.67[c]	.45[d]	.64[a]	.40[b]	.67[c]	.45[d]
Reading	.70	.49	.72	.52	.64	.41	.70	.49
Total reading	.71	.51	.73	.54	.64	.41	.69	.48
	Mathematics							
Computation	.56	.32	.60	.36	.48	.23	.60	.36
Concepts	.61	.37	.67	.45	.55	.30	.68	.46
Problem solving	.57	.32	.65	.42	.55	.30	.66	.44
Total mathematics	.60	.37	.66	.44	.57	.32	.69	.48

[a] Linear r in this column = Pearson correlation coefficient.

[b] Linear r^2 in this column = coefficient of determination.

[c] Quadratic R_Q in this column = Index of Correlation.

[d] Quadratic R_Q^2 in this column = coefficient of determination.

$p < .0001$

Table F

Relationship Between Michigan Educational Assessment Program Scores and Metropolitan Achievement Test Scores for Seventh-Grade Black Students and White Students

Subtest	Black (N = 204)				White (N = 745)			
	Linear		Quadratic		Linear		Quadratic	
	$\underline{r}$	$\underline{r}^2$	R_Q	R_Q^2	$\underline{r}$	$\underline{r}^2$	R_Q	R_Q^2
	Reading							
Word knowledge	.60[a]	.36[b]	.64[c]	.40[d]	.68[a]	.46[b]	.70[c]	.49[d]
Reading	.71	.50	.73	.54	.74	.54	.75	.56
Total reading	.69	.48	.73	.53	.74	.54	.75	.57
	Mathematics							
Computation	.60	.36	.61	.37	.71	.50	.77	.59
Concepts	.57	.32	.60	.37	.75	.56	.79	.62
Problem solving	.56	.32	.59	.35	.72	.52	.78	.61
Total mathematics	.69	.47	.71	.50	.77	.60	.83	.69

[a] Linear $\underline{r}$ in this column = Pearson correlation coefficient.

[b] Linear $\underline{r}^2$ in this column = coefficient of determination.

[c] Quadratic R_Q in this column = Index of Correlation.

[d] Quadratic R_Q^2 in this column = coefficient of determination.

$p < .0001$

Table G

Relationship Between Michigan Educational Assessment Program Scores and Metropolitan Achievement Test Scores for Seventh-Grade Male and Female Students

Subtest	Male (N = 492)				Female (N = 457)			
	Linear		Quadratic		Linear		Quadratic	
	r	r^2	R_Q	R_Q^2	*r*	r^2	R_Q	R_Q^2
Reading								
Word knowledge	.72[a]	.52[b]	.73[c]	.53[d]	.70[a]	.49[b]	.74[c]	.55[d]
Reading	.75	.57	.76	.57	.76	.58	.79	.63
Total reading	.76	.58	.77	.60	.76	.57	.79	.63
Mathematics								
Computation	.74	.55	.78	.61	.71	.51	.77	.59
Concepts	.77	.59	.80	.64	.77	.60	.81	.66
Problem solving	.76	.57	.80	.64	.73	.54	.80	.63
Total mathematics	.81	.65	.84	.71	.80	.63	.85	.72

[a] Linear *r* in this column = Pearson correlation coefficient.

[b] Linear r^2 in this column = coefficient of determination.

[c] Quadratic R_Q in this column = Index of Correlation.

[d] Quadratic R_Q^2 in this column = coefficient of determination.

$p < .0001$

Table H

Relationship Between Michigan Educational Assessment Program Scores and Metropolitan Achievement Test Scores for Seventh-Grade Male and Female Black Students

Subtest	Male (N = 99)				Female (N = 105)			
	Linear		Quadratic		Linear		Quadratic	
	r	r^2	R_Q	R_Q^2	r	r^2	R_Q	R_Q^2
				Reading				
Word knowledge	.67[a]	.45[b]	.69[c]	.48[d]	.63[a]	.39[b]	.67[c]	.44[d]
Reading	.78	.60	.78	.61	.67	.44	.72	.52
Total reading	.76	.57	.77	.59	.69	.47	.74	.54
				Mathematics				
Computation	.65	.43	.65	.43	.53	.28	.56	.31
Concepts	.52	.27	.54	.29	.67	.45	.69	.48
Problem solving	.57	.33	.59	.34	.61	.37	.62	.38
Total mathematics	.68	.46	.69	.47	.73	.53	.75	.56

[a] Linear r in this column = Pearson correlation coefficient.

[b] Linear r^2 in this column = coefficient of determination.

[c] Quadratic R_Q in this column = Index of Correlation.

[d] Quadratic R_Q^2 in this column = coefficient of determination.

$p < .0001$

Table I

Relationship Between Michigan Educational Assessment Program Scores and Metropolitan Achievement Test Scores for Seventh-Grade Male and Female White Students

Subtest	Male (N = 393)				Female (N = 352)			
	Linear		Quadratic		Linear		Quadratic	
	r	r^2	R_Q	R_Q^2	r	r^2	R_Q	R_Q^2
Reading								
Word knowledge	.69[a]	.48[b]	.70[c]	.49[d]	.68[a]	.46[b]	.72[c]	.51[d]
Reading	.72	.52	.72	.52	.75	.57	.78	.61
Total reading	.74	.54	.74	.55	.74	.55	.77	.60
Mathematics								
Computation	.73	.53	.78	.61	.71	.51	.77	.59
Concepts	.78	.60	.80	.64	.75	.56	.79	.62
Problem solving	.75	.57	.80	.63	.71	.50	.79	.62
Total mathematics	.80	.64	.84	.71	.77	.60	.83	.70

[a] Linear r in this column = Pearson correlation coefficient.

[b] Linear r^2 in this column = coefficient of determination.

[c] Quadratic R_Q in this column = Index of Correlation.

[d] Quadratic R_Q^2 in this column = coefficient of determination.

$p < .0001$

APPENDIX B

Figure Captions

Fig. 1.--Scatter diagram of total number of reading objectives mastered (horizontal) and total reading grade equivalent scores (vertical) of fourth-grade students (A = 10, B = 11, C = 12, D = 13, E = 14, F = 15, etc.).

Fig. 2.--Scatter diagram of total number of mathematics objectives mastered (horizontal) and total mathematics grade equivalent scores (vertical) of fourth-grade students (A = 10, B = 11, C = 12, D = 13, E = 14, F = 15, etc.).

Fig. 3.--Scatter diagram of total number of reading objectives mastered (horizontal) and total reading grade equivalent scores (vertical) of seventh-grade students (A = 10, B = 11, C = 12, D = 13, E = 14, etc.).

Fig. 4.--Scatter diagram of total number of mathematics objectives mastered (horizontal) and total mathematics grade equivalent scores (vertical) of seventh-grade students (A = 10, B = 11, C = 12, D = 13, E = 14, etc.).

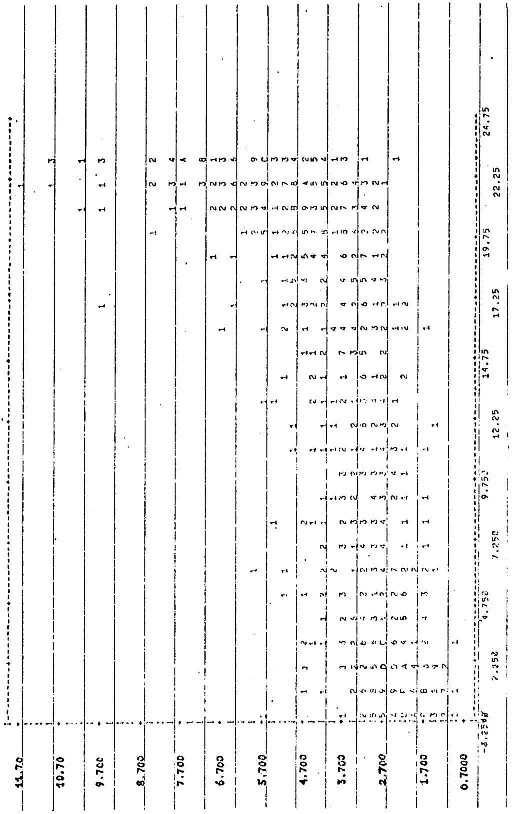
11.70
10.70
9.700
8.700
7.700
6.700
5.700
4.700
3.700
2.700
1.700
0.7000
-0.2500
2.250
4.750
7.250
9.750
12.25
14.75
17.25
19.75
22.25
24.75

Figure 1

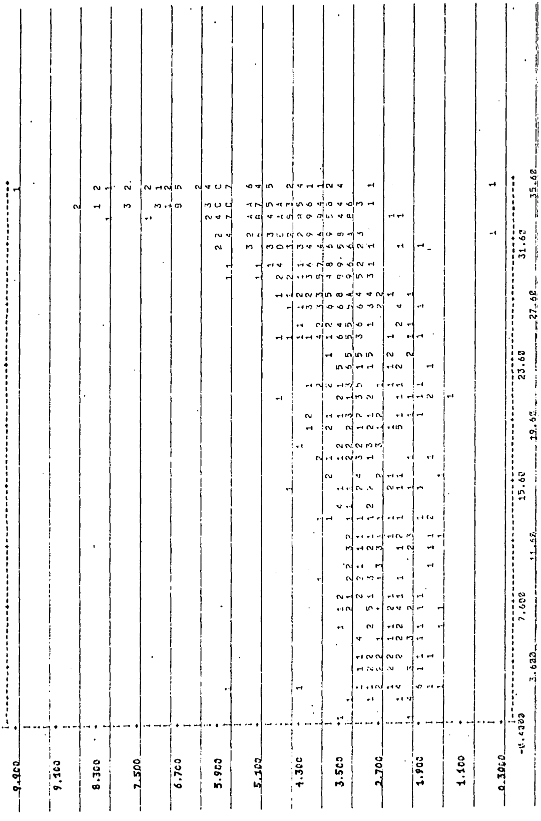
9.900
9.100
8.300
7.500
6.700
5.900
5.100
4.300
3.500
2.700
1.900
1.100
0.3000
-0.4000
3.600
7.600
11.60
15.60
19.60
23.60
27.60
31.60
35.60

Figure 2

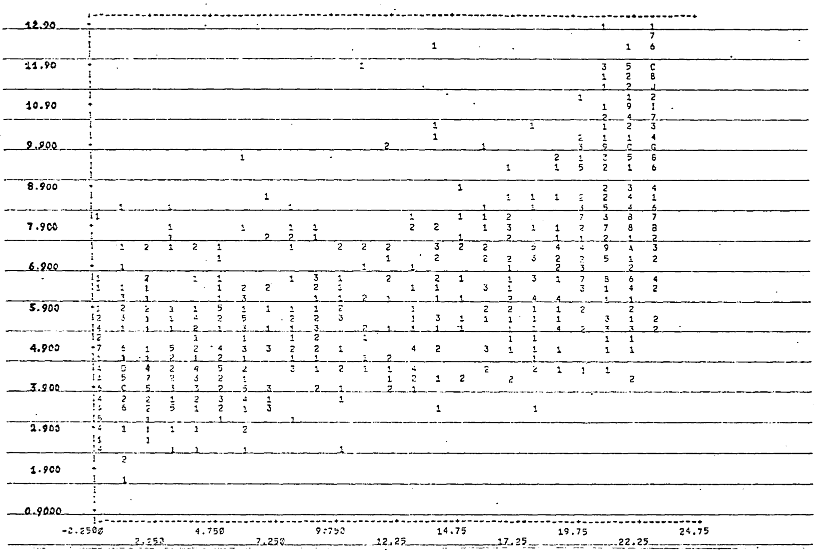
12.90
11.90
10.90
9.900
8.900
7.900
6.900
5.900
4.900
3.900
2.900
1.900
0.9000
-2.2500
2.250
4.750
7.250
9.750
12.25
14.75
17.25
19.75
22.25
24.75

Figure 3

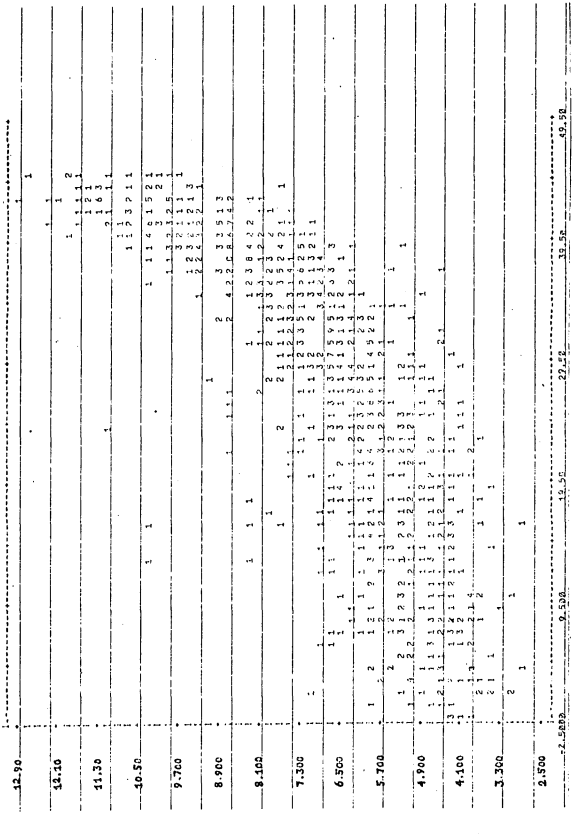
12.90
12.10
11.30
10.50
9.700
8.900
8.100
7.300
6.500
5.700
4.900
4.100
3.300
2.500
-2.500
9.500
19.50
29.50
39.50
49.50

Figure 4