INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

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

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

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

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.

University Microfilms International
300 N. Zeeb Road
Ann Arbor, MI 48106

Schaefer, Mary Miller

A COMPARISON OF RELIABILITY ESTIMATES FROM SINGLE AND DOUBLE ADMINISTRATIONS OF CRITERION-REFERENCED TESTS

City University of New York PH.D. 1983

A COMPARISON OF RELIABILITY ESTIMATES FROM SINGLE AND DOUBLE ADMINISTRATIONS OF CRITERION-REFERENCED TESTS

by

MARY MILLER SCHAEFER

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

1983

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

4/25/83
date

Chairman of Examining Committee

4-25-83
date

Executive Officer

Dr. Alan L. Gross

Dr. Max Weiner

Dr. David Rindskopf
Supervisory Committee

The City University of New York

Abstract

A COMPARISON OF RELIABILITY ESTIMATES FROM SINGLE AND DOUBLE ADMINISTRATIONS OF CRITERION-REFERENCED TESTS

by

Mary Miller Schaefer

Adviser: Professor Alan Gross

The primary purpose of individualized instructional programs is to maximize each student's opportunity to learn. Measurement of students in an individualized instructional program is keyed to the instruction provided to each student, and the appropriate measurement information is how far the student has progressed along the instructional continuum. The student's achievement is compared to a criterion and thus the appropriate type of measurement is a criterion-referenced test. A criterion-referenced test is one which is designed to evaluate whether a student has met the acceptable performance standard(s) of instructional objectives in an instructional program. Those who meet or exceed performance standards are considered masters while those who do not meet the performance standards are nonmasters.

The determination of reliability for criterion-referenced tests has centered on the view of reliability as the consistency of mastery classification decisions from one testing to another or on two forms of a criterion-referenced test. Three models for determining reliability

were examined in a test-retest setting to investigate the effect of five student and test characteristics on the resulting reliability coefficients. One model (k), considered the standard, used data from two test administrations. The two other models ($\hat{k}_H$ and p_{cs}) were developed for use when only data from a single test administration were available. The single administration estimates were compared to the standard for each testing condition and all the criterion-referenced reliability coefficients were compared to norm-referenced reliability coefficients computed on the same data.

Coefficient k generally has the lowest values and the largest standard errors. The size of k increases with increased test length, decreases when a cut-off score is set at an extreme point such as 100 percent, and seems to be maximized by larger sample sizes which also results in smaller standard errors. The heterogeneity of test content had a mixed effect on k.

The estimate of kappa, $\hat{k}_H$, generally had slightly larger mean values than the mean values of k and smaller standard errors. The mean values of $\hat{k}_H$ increased with test length, and the mean values were highest for a cut-off score of 80 percent. The change in sample size did not predictably affect $\hat{k}_H$ and the mean values of $\hat{k}_H$ were highest for the low ability groups. The violation of test item homogeneity appeared to reduce $\hat{k}_H$.

The coefficient of agreement p_{cs} was the highest reliability coefficient with the smallest standard errors, across all analyses. The high values of p_{cs} were partly due to the difference in scaling between p_{cs} and the kappa coefficients. The mean values of p_{cs} did increase with test length and the maximum mean values occurred at the

60 percent cut-off score. The size of the p_{cs} coefficient varied for ability level by cut-off score and test length. The p_{cs} coefficients' mean values were higher for one classroom samples than for the larger group, and the heterogeneity of test content increased the size of the mean values of p_{cs}.

The estimate of k, k_H, overestimated k under all conditions except heterogeneous test content. There was little relation between p_{cs} and k under any conditions, with p_{cs} being consistently much larger than k.

Additionally there was no consistent pattern of relationships between the norm-referenced and criterion-referenced coefficients.

ACKNOWLEDGEMENTS

I would like to acknowledge the members of my dissertation committee: Dr. Alan Gross, my advisor; Dr. David Rindskopf and Dr. Max Weiner.

I also wish to acknowledge the following people without whom the dissertation could not have been done:

Dr. Susan Gross, Montgomery County Public Schools provided guidance, inspiration and a place to work from the first draft of the proposal through the final copy.

Suzette Brown, Joseph Houston and George Roberts provided expert computer programming of complicated problems.

Dale Conlan did an excellent job typing the initial proposal and all the drafts of the final document.

Dr. Steve Frankel and Dr. Joy Frechtling, Montgomery County Public Schools, gave me a job on a project which led to the idea and the data for this dissertation.

Finally and most importantly my husband Dr. Ernst J. Schaefer gave me unwavering support and encouragement. Without him I would be A.B.D.

I dedicate this dissertation to my children Caroline, Christopher, and Peter.

TABLE OF CONTENTS

TABLE OF CONTENTS - (Continued)

TABLE OF CONTENTS - (Continued)

LIST OF TABLES

LIST OF TABLES - (Continued)

LIST OF TABLES - (Continued)

LIST OF TABLES - (Continued)

LIST OF TABLES - (Continued)

LIST OF TABLES - (Continued)

LIST OF TABLES - (Continued)

LIST OF FIGURES

Chapter 1

INTRODUCTION

Background

In many of the instructional programs which are based on a mastery learning strategy, there is an extensive use of tests to determine when a student has met an acceptable level of performance or criterion. The information provided by the test is used to evaluate the student's mastery of objectives for the purpose of assigning him/her to the next appropriate level of instruction. Typically, a mastery score (or cut-off score) is set on each subset of test items to allow the teachers to assign students into one of two mutually exclusive categories--masters and non-masters--based on their performance on the items measuring an objective (Hambleton, 1974). Mastery, then, is used as a label which characterizes an individual's achievement with respect to the instructional objectives (Meskauskas, 1976). Not only have individualized or objectives-based instructional programs been developed by such large groups as the American Institutes for Research and the Westinghouse Learning Corporation (Program for Learning in Accordance with Needs--PLAN), but also by county and city school systems which have been attempting to make instruction more responsive to students' needs. Once an instructional program and the appropriate tests have been developed, program designers and implementers are faced with

the problem of determining how much a student has learned. This determination should necessarily encompass questions of student achievement and also questions about the reliability and validity of the tests which are typically used to determine this achievement.

If tests are used in an objectives-based program to assign students to a mastery state with respect to particular objectives, it is important to determine whether the tests are reliable. It has been shown that classical reliability estimates are generally considered inappropriate for criterion-referenced measures and a considerable amount of literature has been published which deals with the problem of reliability for criterion-referenced tests. However, as will be illustrated in Chapter 2, the literature is far from consistent in the way the authors conceptualize the concept of reliability for criterion-referenced tests.

Concepts of Reliability

Hambleton, Swaminathan, Algina and Coulson (1978) delineate three concepts of reliability that arise in the context of criterion-referenced testing: reliability of mastery classification decisions, reliability of criterion-referenced test scores, and reliability of domain score estimates (pp. 15-23).

The first concept, the reliability of mastery classification decisions, is concerned with the consistency with which individuals are classified as masters or nonmasters on two forms of a criterion-referenced test. Masters are individuals who meet or exceed the standard of performance (or criterion) while nonmasters are those whose achievement does not equal the criterion level of performance.

The second type of reliability, the reliability of criterion-referenced test scores, refers to the stability of deviations from the criterion score on two forms of a criterion-referenced test. For example, a student's score of 24 (out of a criterion score of 25) on two forms of a criterion-referenced test would be perfect reliability because the deviation of -1 is the same across both forms. Scores of 24 on Form 1 and 20 on Form 2 would indicate lower reliability because the score deviates -1 from the criterion score on one form and -5 from the criterion on the second form.

The third type of reliability, the reliability of domain score estimates, is different from the first two types in that it does not involve setting a criterion score and thus does not determine masters and nonmasters. This reliability concerns the consistency of students' scores across two forms of parallel tests. The reliability of domain score estimates is appropriate when the purpose of the test is to estimate the number or the proportion of similar items which students can answer correctly.

Support for reliability of mastery classification decisions has been advanced primarily by Hambleton and Novick (1973) and Swaminathan, Hambleton, and Algina (1974), who maintain that the crucial reliability issue is whether or not an examinee is consistently assigned to the same side of the criterion on two tests--i.e., is the student consistently found to be a master or a nonmaster? Swaminathan, Hambleton and Algina (1974) recommend the use of coefficient kappa (Cohen, 1960) to determine reliability for criterion-referenced tests.

Two other models for determining reliability which are based on this definition of criterion-referenced reliability are those developed by Huynh (1976) and Subkoviak (1976). Huynh's (1976) model can be used to estimate reliability from a double or a single administration of a criterion-referenced test, and Subkoviak's model is intended for use when there has been only a single administration of a criterion- referenced test. Huynh (1976) calls the primary purpose of criterion-referenced testing the classification of examinees into the two achievement states of mastery and nonmastery. Subkoviak (1976) defines a coefficient of agreement for a student as the probability that the student is assigned to the same mastery state on two parallel tests.

The Problem

In an instructional program which is based on a mastery learning strategy, the object is for all students to reach mastery by providing instruction in such a way that students can proceed at their own pace and so they can master essential skills upon which subsequent learning will be built. The philosophy is that all students can reach mastery if certain conditions are met to account for individual differences in learning. One issue which is part of a mastery learning strategy is whether the classification of a student as a master is equivalent to labelling the student proficient in that particular learning area. In order to have confidence in the classification of a student as a master, it is necessary to have tests which are reliable. But there remains a question about whether there is one method of estimating reliability

that is most appropriate for accurately determining that a student is a master and is thus proficient in a unit of instruction.

Additionally, program developers are increasingly specifying program objectives which are formulated in terms of absolute standards. These standards are used to judge the significance of educational programs. For example, school districts which have instituted or plan to institute competency tests for students set absolute standards which the students must meet in order to be determined to be competent. In this type of assessment the issue of reliability really concerns the consistency with which examinees are assigned to mastery or nonmastery categories--that is, the reliability of mastery classification decisions.

The focus of the present research will be on three reliability measures which reflect the same theoretical view of reliability for criterion-referenced assessment--the reliability of mastery classification decisions. This view of reliability is relevant to current educational practitioners and it is expected that the present research may point out particular sensitivities of the different estimates which may be useful for deciding which estimates to use under specific conditions.

One purpose of this research, then, will be to examine three reliability estimates coefficient kappa (k) (Cohen, 1960) and suggested for use with criterion-referenced tests by (Swaminathan, Hambleton, Algina (1974), Huynh's (1976) single administration estimate of coefficient kappa ($\hat{k}_H$) and Subkoviak's (1976) single administration estimate) using the same data set for each estimate.

The computation of coefficient kappa is only possible when there have been two test administrations whereas the other two coefficients can be computed when there has been only one test administration. Student outcomes on criterion-referenced tests will be examined in a test and retest setting to determine the sensitivity of the three reliability estimates to an entire set of test and examinee characteristics.

A second purpose of this research will be to try to determine the relationship of the two estimates of reliability to kappa. Coefficient k is an ideal measure because it requires two test administrations, and when two test administrations are possible, k represents a standard. Thus, if possible, k is the reliability coefficient one would choose to compute. However, it is not always feasible to administer a test twice and in those instances one needs to estimate reliability using only a single test administration. If a reliability estimate correlates positively with k, one can consider that the estimate behaves in the same way under particular conditions. Therefore, correlation coefficients will be computed between each of the two reliability estimates and k under varying conditions to investigate how the relationship of these two measures to k changes. Further, in order to investigate whether k is a good estimate of k under different conditions the mean difference score $(\overline{k-\hat{k}})$ will be computed for the conditions which are varied. This analysis will not be undertaken for Subkoviak's estimate of reliability since it is not intended as an estimate of k.

Another issue concerns the relationship of reliability to (1) cut-off score and (2) test length. In other words, how do each of the three estimates behave as test length and cut-off score - i.e., mastery rates - vary, and further how does the relationship of the two measures to k change when mastery rates vary?

Another question is how does (3) variability in students' ability affect the reliability estimates? If a test is administered to students who are more proficient in the subject area, as determined by a standardized achievement test in mathematics, as well as to students who are less proficient in the subject area, how does the relationship between the two estimates and k change for these two groups of students?

Two other issues concern (4) the size of the sample and (5) the heterogeneity of test content, each of which could have an effect on the reliability estimates. Heterogeneous test content is achieved through combining items from different levels of objectives within the skill areas of multiplication or division. The objectives in the mathematics program are hierarchical and an example of heterogeneous test content would be items from several levels combined into one test. These items might include: multiplication of one-digit numbers by one-digit numbers under five; multiplication of two-digit numbers by 6, 7, 8 and 9; multiplication of two-digit numbers by two-digit numbers.

The preceding five characteristics were selected for analysis because they are characteristics which would probably affect reliability estimates for norm-referenced tests and it is thus of

interest to investigate how these characteristics affect reliability estimates for criterion-referenced tests.

These major purposes of the research may be summarized as follows:

I. How is coefficient kappa, k, the standard affected by test length (n), cut-off score, student ability, sample size (N), and test content heterogeneity?

II. How are Huynh's single sample estimate $\hat{k}_H$ and Subkoviak's coefficient $p_c(s)$ affected by test length (n), cut-off score, student ability, sample size (N) and test content heterogeneity?

III. Do $\hat{k}_H$ and $p_c(s)$ behave in the same way as k for these differing conditions?

IV. Is $\hat{k}_H$ a good estimate of k for these varying conditions?

A final issue is whether a reliable determination of a student's mastery of an area can be considered similar to student proficiency. If an outside criterion of proficiency can be identified, one can assess the validity of a criterion-referenced test, i.e., the accuracy with which it distinguishes between masters and nonmasters. It is then of interest to investigate the relationship between this validity and the estimated reliability of the test.

Chapter 2

REVIEW OF THE LITERATURE

The necessity of developing individualized instructional programs has become apparent from the results of research during the past decade or two. Several studies have shown that students differ along many dimensions such as interests, motivation, and learning rate, and thus instruction aimed at a group of students may be inappropriate for individual students. Instructional programs which are individualized have as their primary purpose to maximize each student's opportunity to learn. Therefore, measurement of students in an individualized instructional program is keyed to the instruction provided to each student, and the appropriate information gathered by measurement is how far the student has progressed along the instructional continuum. The student's achievement is compared to a criterion and thus the appropriate type of measurement is a criterion-referenced test. A criterion-referenced test is one which is designed to evaluate whether a student has met the acceptable performance standard(s) of instructional objectives in an instructional program. Criterion-referenced tests are generally used quite frequently in individualized instructional programs because the student's mastery or lack of mastery of the objectives being measured determines his placement for further instruction. In an individualized instructional program, one of the crucial assumptions is that all students can attain mastery of all the objectives, but the amount of time needed to attain mastery may vary among students.

While criterion-referenced measurement is the focus of the present research, a discussion of criterion-referenced measurement cannot ignore the concept of mastery learning which had its historical roots in the 1920's and which became an important concept again in the 1960's. The first introduction of the mastery learning concept in this country's schools occurred in the Winnetka Plan which was the work of Washburne (1922). Block (1971) has hypothesized that while the program flourished in the 1920's, it was probably abandoned because of a lack of technology to sustain it. The concept surfaced again in the early 1960's with the movement to develop programmed instruction. The purposes of programmed instruction were to facilitate student learning by breaking down even the most complex skills into simpler components which the student could learn and to give the student immediate feedback on his/her response to questions about the behavior(s) learned in progressing through the instructional sequence. Students were presented with material in an "instructional frame" and at the completion of the frame, they were asked a question to determine if they had adequately learned the material. If the response were correct, the student's learning was reinforced and he/she proceeded to the next instructional frame. If, however, the response were incorrect, the error was corrected immediately. Each student progressed through the instructional sequence at his/her own rate. Many large scale educational systems were developed (Project PLAN, IPI) in the 1960's which were attempts to improve on programmed

instruction and which were based on a mastery learning paradigm developed by Bloom (1968).

Bloom's (1968) mastery learning model grew out of an attempt to improve programmed instruction because not all types of students were learning effectively in a programmed instruction situation. The units available did not adequately handle individual differences in learning, and Bloom thus elaborated on a conceptual model of school learning developed by Carroll (1963) to develop a mastery learning model.

Carroll (1963) proposed a model of school learning which had five variables: aptitude, quality of instruction, ability to understand instruction, perseverance, and time allowed for learning. These five variables interacted to determine degree of learning for each student.

Aptitude is viewed by Carroll (1963) as the amount of time required by each learner to achieve mastery of a learning task. This viewpoint would assume then that given an adequate amount of time any learner would be able to achieve mastery of the learning task.

The second variable in this school learning model is quality of instruction, and in this model, quality of instruction is defined in terms of individual learners. Instead of assuming that there be quality of instruction for an entire group or classroom of learners, Carroll (1963) defines quality of instruction as the degree to which presentation, explanation, and ordering of the learning task is the best for each student.

The Ability to understand instruction is also defined in terms of the individual learner and is the ability of the learner to understand the nature of the learning task and how he is to go about learning it.

Perseverance is the amount of time which the student is willing to spend actively involved in learning a task. Closely related to perseverance is the time allowed for learning, the fifth variable in Carroll's (1963) model. This variable is simply the amount of time allocated by the teacher to each student for a given learning task.

In developing his strategy for mastery learning from Carroll's (1963) model, Bloom (1968) first discussed the relationship of aptitude and achievement which he felt was implicit in Carroll's model. That is, if learners are normally distributed with respect to aptitude for a learning task and all students are then given the same amount and quality of instruction with the same time allowed for learning, then the learners will be normally distributed with respect to achievement. However, if the amount and quality of instruction and learning time allowed for instruction are varied for each student, then even if the students are normally distributed with respect to achievement, the majority of students will achieve mastery. The correlation between aptitude and achievement should be very low if the instructional variables are made appropriate for each learner.

Bloom (1968) incorporated the major variables in Carroll's model into a suggested strategy for mastery learning. He identified some preconditions, suggested some necessary operating procedures and described possible measurable outcomes.

To begin with, one required precondition is the "specification of the objectives and content of instruction and the translation of these specifications into summative evaluation procedures" (Bloom, 1968, p. 57). Mastery must be defined and one must be able to determine if students have achieved mastery.

Useful operating procedures consist of both the division of the subject to be taught into smaller units and the use of formative evaluation instruments to measure mastery of learning tasks. The division of a subject into smaller units can be accomplished by using the ideas of Bloom (1956) and Gagne (1965) to break down each subject into specific elements. The formative evaluation instruments are brief diagnostic tests based on particular objectives and are used to determine mastery or nonmastery of objectives. The tests are diagnostic in the sense that achievement on the tests determines the next step for the students in their instructional sequence. If the students do not master the tests, they can be assigned to remediation activities, and if they do master, they continue to the next scheduled learning task. The results of these tests ideally provide positive feedback to students who master them and provide prescriptive feedback to students who do not master them. Additionally, the tests provide feedback to the teachers about particular instructional aspects of the unit which may need modification. These formative tests are keyed to a particular objective, and a standard (or criterion) is set on the tests which a student must meet to achieve mastery. The students' achievement is compared to the standard, not to each other's achievement, and thus the tests are often referred to as criterion-referenced tests.

An achievement test's scores provide two types of information. One type is the degree to which the student has achieved the criterion performance. The second type of information provided by an achievement test score is the ordering of students according to their performance on the test. The first type of information - the student's attainment of a standard or criterion is criterion-referenced. A test which yields that type of information is a criterion-referenced test. The second type of information - the student's ranking in a group - is norm-referenced. A test which is designed to yield that type of information is a norm-referenced test.

One of the first articles to discuss and to differentiate between these two types of measurements was written by Glaser (1963). Glaser's (1963) initial concern was with the determination of the effectiveness of programmed learning and teaching machines. In this context he proposed the distinction between the two types of measurement information which can result from achievement testing. The first type of measurement is the information about whether a student has attained a certain performance level on the test and the second piece of information concerns where a student ranks among other students who have taken the same test. A student's attainment of a performance level is considered criterion-referenced while a student's capability in comparison with other students is considered norm-referenced. Glaser pointed out that when students are assessed in relation to a criterion or standard, one obtains information about students' competence which is independent of their relation to the performance of others. It is this type of information which he

felt was most important when determining the effectiveness of instructional technology.

A further point made by Glaser is the importance of the difference in the selection criteria of test items depending upon the type of measurement to be used. For an achievement test of individual differences (i.e., norm-referenced) one would choose items which will discriminate among individuals having had the same treatment while for a test to distinguish between groups one would choose items to show that a particular treatment was effective. Glaser concluded the article by mentioning a growing recognition in the field of education that the assessment of levels of competence needed many new considerations different from those which grew out of the most wide-spread type of assessment - norm-referenced assessment. This article really called for new approaches to assessing learning outcomes.

It was not until the late 1960's and early 1970's that educational testing experts really became publicly concerned with the differences between criterion-referenced testing and norm-referenced testing. In a paper prepared for Educational Measurement, Glaser and Nitko (1970) specified:

> A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards (p. 653).

Glaser went on to explain that these types of tests were to be used to describe an individual's performance in relation to a "specified domain of tasks." Therefore, the performance standard(s) must be determined before the test is constructed. Since one is interested in assessing a student's achievement of performance standards, tests

which are constructed for this purpose can be interpreted without referring to a norm-group, and this fact is a crucial distinction between criterion-referenced and norm-referenced tests.

Popham and Husek (1969) enumerated six points which differentiate criterion-referenced tests from norm-referenced tests, and these six points elaborated Glaser's initial distinction between these two types of tests. The first issue is that of variability. For norm-referenced tests variability is a desirable property as one is interested in the comparison of scores and maximizing individual differences. One wants to be able to discuss the position of a student's score in relation to the other scores on the test and, therefore, variability is important. For criterion-referenced tests, however, the score's meaning results from the relation of the score to the criterion score or performance standard. Variability of scores for criterion-referenced tests is not desirable since the object is to have <u>all</u> students achieve mastery.

A second area of difference between norm-referenced and criterion-referenced tests is that of item construction. For norm-referenced tests, one is interested in differentiating among students who are taking the test. Therefore, item choices are made so that items are not too easy or too hard and so that they ideally maximize the differences among the different achievement levels of students. For a criterion-referenced test, however, the aim of the item-writer is to insure that each item is an accurate reflection of the performance standard one is measuring. It is not important if the item is difficult or easy, or whether the item discriminates

among students as long as the item is representative of the behavior which is considered criterion behavior.

As more attention was being directed toward criterion-referenced measurement, Ebel (1971) and Block (1971) defined criterion-referenced measurement and discussed its limitations (Ebel) and its potential (Block). Ebel maintained that the difference between norm-referenced and criterion-referenced measurements was "in the quantitative scales used to express how much an individual can do." (p. 283). For norm-referenced tests, Ebel asserted that the scale is "anchored in the middle on an average level of performance for a particular group of individuals" while for criterion-referenced measurements, "the scale is usually anchored at the extremities a score at the top indicating complete or perfect mastery of some defined abilities; one at the bottom indicating complete absence of these abilities." (p.283). Ebel continued to delineate what he felt were limitations for criterion-referenced measurement: inability to tell all one needs to know about educational achievement, difficulty of obtaining good criterion-referenced measures, and lack of meaningful criteria for criterion-referenced measures. Block (1971) in a companion article rebutted Ebel's limitations and also talked about the distinction between norm-referenced and criterion-referenced measures. Block maintained that scale properties alone could not distinguish the two types of measures and that the distinction could only be made by examining . . . "the purposes for which they are made, the manner in which they are obtained, the specificity of information they provide regarding student learning and the purposes for which they are used . . . ." (p. 290).

Hambleton and Novick (1973) distinguished between norm-referenced and criterion-referenced tests in terms of the kinds of decisions they are designed to make. They characterized testing as a "decision-theoretic" process. Norm-referenced measurement is seen by them as particularly appropriate when one is interested in "fixed quota selection" or in "ranking individuals on some ability continuum." (p. 162). Criterion-referenced measurement, however, is most appropriate for "quota-free selection" where there is no restriction on students who can "exceed the cut-off score or threshold on a criterion-referenced test." (p. 163). The purpose of the test is to divide students into two groups which are mutually exclusive, masters and nonmasters.

Millman (1974) coined a new term for criterion-referenced measurements - domain-referenced tests. His definition of a domain-referenced test (DRT) was:

> Any test consisting of a random or stratified random sample of items selected from a well-defined set or class of tasks (a domain). (p. 315).

Millman asserted that since a DRT was made up of a sample of items from a well-defined population of items, one could estimate an "examinee's domain score or level of functioning, defined as the percent of the population of items the examinee could answer correctly or in a given direction." (p. 315).

In a relatively recent review of developments in the field of criterion-referenced testing and measurement, Hambleton, Swaminathan, Algina, and Coulson (1978) differentiate among criterion-referenced tests, domain-referenced tests, and objectives-referenced tests. Using a definition of Popham's (1975):

> A criterion-referenced test is used to ascertain an individual's status (referred to as a domain score) with respect to a well-defined behavior domain. (p. 130).

they see relatively no difference between criterion-referenced and domain-referenced tests. However, unlike criterion-referenced tests, objectives-referenced tests have no domain of behavior specified and items are not considered to be representative of any behavior domain." (p. 3).

It is clear from the preceding discussions that there is disagreement in the literature about the meaning and the utility of criterion-referenced measurement. In fact, the 1978 Annual Meeting of the American Educational Research Association replaced its presidential address by a presidential debate: The Case for Norm-Referenced Measurements (Ebel, 1978) vs. the Case for Criterion-Referenced Measurements (Popham, 1978). One of the problems implicit in the division of theoretical experts pro or con criterion-referenced testing is the limited development of procedures for dealing with statistical and psychometric questions.

While reliability and methods for determining reliability of criterion-referenced tests will be dealt with in detail in Chapter 3, briefly it can be noted that classical estimates of reliability depend upon test score variability. Therefore, these estimates are not appropriate for use with criterion-referenced tests since there may be instances when all students achieve perfect scores and there would be no test score variability. The lack of test score variability would also render classical stability estimates questionable. The problem with validity estimates is the same.

Many estimates of validity are based on correlations which, again, will approach zero when criterion-referenced measures are correlated with normally distributed variables. For criterion-referenced tests, judgment of a test's relevance to the performance standard may be considered a more logical means of determining the test's validity.

The traditional use of item analysis in test validation has been to find items which do not discriminate among groups of students, generally between the more and less knowledgeable students. Items are discarded if they are too difficult, too easy, or ambiguous. However, these judgments do not apply for criterion-referenced tests. An item which discriminates positively (more of the more knowledgeable students answer the question correctly) would be judged as adequate, and an item which discriminates negatively (more of the less knowledgeable students answer the question correctly) should probably be discarded. However, items which do not discriminate at all may be included on the test if they represent essential knowledge which all students must acquire.

A final distinction between norm-referenced and criterion-referenced tests arises from the reporting and interpretation of the test scores. For an achievement test, scores are generally reported as standard scores, grade equivalents, stanines, all of which are scores which display an individual's performance relative to the group of test takers. For a criterion-referenced test, however, the relevant information is whether or not an individual has reached a performance standard (mastery). Therefore, it might be most appropriate to report simply success or failure in reaching the

performance standard. In some instances, degrees of failure (i.e., how close did the student come to success) may be reported depending upon the use to be made of the data.

However, since Glaser's (1963) seminal article on criterion-referenced testing, there have been major developments in the field of criterion-referenced testing of procedures for dealing with statistical and psychometric questions. It may be noted, however, that many of the procedures have been theoretical only, and many have not been tried out with actual criterion-referenced data. Also, many procedures make certain assumptions about the distribution of the data, and there are questions about what kinds of results would occur if the model's assumptions were violated. In general, despite a relatively rapid growth in psychometric techniques for dealing with criterion-referenced measurements, many questions remain. One of the areas which has a lack of answers to many questions is the area of reliability.

The necessity for developing models for determining reliability of criterion-referenced tests can be introduced by a brief review of classical reliability theory. Appropriate reliability measures for any instrument consist of test-retest reliability (stability of test performance over time), internal consistency, and parallel forms reliability (consistency of performance on parallel tests).

Underlying the determination of reliability measures in classical test theory is the concept of true score. In fact, reliability is defined as the ratio of true score variance to observed score variance or as the squared correlation between true scores and observed scores (Lord and Novick, 1968). However, since true scores are

unobservable, one must estimate them, and in classical test theory there are two ways to estimate true scores. One method requires that either the same test be administered at two different times or that parallel forms be administered at two different times. These testing situations would enable one to determine test-retest or parallel forms reliability. With one test administration, one can determine the internal consistency of a test by computing split-half test score correlations or by computing the correlation of item total and test score. These methods for determining reliability for a norm-referenced test are based on test score variance. Large reliability coefficients suggest that a student's performance will be approximately the same on two test administrations, and they also indicate that the test discriminates among the students. Discrimination among students is a major consideration for norm-referenced tests.

However, as we have seen earlier, the goal of criterion-referenced tests is not to maximally discriminate among students. The goal of a criterion-referenced test is to determine the achievement (or mastery) of a performance standard. The concern is simply whether students do or do not reach a performance standard. Those students who do reach this standard are considered masters while those who do not reach the standard are considered nonmasters. The results of a criterion-referenced test would be used to classify students into these two mastery categories. It is possible, and in many cases it is desirable, for all the students who are classified as masters to have achieved a perfect score on the criterion-referenced test. If this is the case, there is no test score

variance for this group of students. Therefore, if one were to use classical reliability methods to investigate the internal consistency of this criterion-referenced test, the reliability coefficient would be close to zero. Indeed, the central issue for determining reliability for criterion-referenced tests is the replicability of the mastery decision, i.e., to assign appropriate students to mastery and nonmastery categories.

The development of models for determining reliability for criterion-referenced tests began very soon after the publication of Glaser's (1963) article. Cox and Graham (1966) developed a coefficient of reproducibility in working with what they felt was a special type of criterion-referenced measure, a sequentially scaled achievement test. This type of test would be constructed so that a student could answer all items up to a certain point - that student's level of attainment - and then would be unable to answer any questions beyond that point. An investigation of a student's score on this type of test would enable one to learn a student's response pattern because of the Guttman scale quality of the test. An analysis of a group of such scores would yield a coefficient of reproducibility. This coefficient would indicate how well a student's response pattern could be reproduced from knowing his total score. Cox and Graham suggest that this coefficient might be used as a "type of reliability estimate across all individuals taking the test." (p. 148).

Popham and Husek (1969) historically seem to be the next to discuss the issue of criterion-referenced test reliability. Yet their discussion consisted primarily of an explication of the

inappropriateness of classical reliability indices for criterion-referenced tests due to the lack of test score variability. They did not develop any alternative models for determining reliability.

Carver (1970) suggested two possibilities for determining the reliability of criterion-referenced tests. One method consisted of administering the same criterion-referenced test to two groups of students at the same level of instruction. A comparison of the percentage of students classified as masters was one indication of the test's reliability. An alternative procedure consisted of the administration of two parallel tests to the same students and a comparison of students classified as masters on the two forms was suggested as the index of reliability. For both procedures, the comparability of percentages of master classifications determined the reliability i.e., the more comparable the percentages, the more reliable the tests.

Livingston (1972a) maintained that it was not necessary to discard concepts of classical test theory to determine the reliability of criterion-referenced tests. The assumption underlying the development of his reliability estimate was that the purpose of a criterion-referenced test was to discriminate each examinee's estimated domain score from a cut-off score. Using this assumption, Livingston went on to replace variance for a criterion-referenced test by the mean squared deviation of the scores from the criterion or cut-off score, rather than from the

mean of sample scores as in classical test theory. Thus, when the criterion score is equal to the mean score, the reliability estimate which Livingston develops is the same as the classical estimate.

Livingston's coefficient drew criticism from several sources. Harris (1972) showed that even though Livingston's coefficient is larger than the classical coefficient, the standard error of measurement is the same and, therefore, the "larger coefficient does not imply a more dependable determination of whether or not a true score falls below or exceeds a given criterion value." (p. 28). Livingston (1972b) countered that reliability is a characteristic of a group of scores, not a single score and that the larger coefficient

> "does imply a more dependable overall determination of whether each true score falls above or below the criterion level when this decision is to be made for every individual score in the distribution." (p. 31).

Shavelson, Block, and Ravitch (1972) criticized Livingston on several issues, but one of the issues was that of the necessity to divide a criterion-referenced test into subscales (other authors subsequently raise this issue as well). They argue that criterion-referenced tests should be "divided into subscales with a criterion for each subscale and that reliability should be estimated for each subscale." (p. 135). They find it inappropriate to report the reliability of test scores which have been obtained by summing across test items which have been written to assess different objectives.

Hambleton and Novick (1973) maintained that the purpose of criterion-referenced tests was the assignment of students to mastery states, and that accordingly, questions of reliability should deal

with whether or not students were consistently assigned to mastery states across parallel forms or retest administrations. Their view then of reliability for criterion-referenced tests is reliability of mastery classification decisions. They propose an index of reliability (p_o) when individuals are to be classified into $\underline{m}$ mastery states:

$$p_o = \sum_{k=1}^{m} p_{kk}$$

In this index, p_{kk} is the proportion of all examinees classified in the kth mastery state on the two test administrations. Then p_o is the observed proportion of decisions which are in agreement. Hambleton and Novick's conceptualization is based on a view that a threshold loss function (as opposed to a squared error loss function such as that of Livingston (1972a)) is a more appropriate way to regard reliability estimates for criterion-referenced tests. The losses which are important are due to misclassification of students to mastery states. The size of the test scores themselves are not important. One might then call the Hambleton and Novick approach to reliability a decision-theoretic approach.

Swaminathan, Hambleton, and Algina (1974) following a decision-theoretic approach expanded upon the index of reliability (p_o) proposed by Hambleton and Novick (1973). They stated the primary purpose of criterion-referenced tests as the assignment of examinees to one of $\underline{k}$ mastery states for each objective measured by items on the test. Thus they defined reliability as the consistency of decisions about mastery states. They felt that the index p_o

did not take into account the proportion of agreement that occurs by chance alone and recommended the use of coefficient k (Cohen, 1960) as an index of reliability. In the case of two administrations of a criterion-referenced test to a group of examinees, p_{ij} = the proportion of examinees placed in the ith mastery state on the first test administration and in the jth mastery state on the second test administration. Coefficient k is defined as:

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

where p_o, the observed proportion of agreement, is given by:

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

and p_c, the expected proportion of agreements, is given by:

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

$p_{i.}$ and $p_{.i}$ represent the proportion of examinees assigned to mastery state i on the first and second test administrations. k can then be considered the proportion of agreement which exists over and above that which can be expected by chance alone. It is interesting to note that the authors "stress" that this coefficient assesses reliability only for tests measuring one objective and that if there are several objectives being measured, there will be several reliability coefficients.

Millman (1974) suggested three different approaches for data analysis for estimating reliability. The first approach, consistency of scores on parallel tests, requires computing a total score for each student on each of two sets of items drawn from the

same domain. The smaller the discrepancy between the two scores, the more reliable the test. The consistency of decisions made from parallel tests is the second approach for which one computes the agreement of decisions suggested by scores on parallel tests. The third approach, consistency of item scores, is determined by examining the consistency of responses to matched items on parallel tests. An example of an index measuring this would be the proportion of all times the test takers passed both or failed both of each pair of matched items. While Millman does mention this third approach, he goes on to suggest that it may actually provide more information about properties of the items than the reliability of the test.

One characteristic of most of the preceding models is the fact that they require two test administrations. In most test applications, it is impractical to assume that two test administrations will exist. Huynh (1976) developed a procedure for estimating coefficient kappa (k) on the basis of a single test administration. Two conditions are necessary for this procedure: (1) underlying true scores on the test are distributed as a beta distribution (a continuous distribution with parameters r and n) and (2) the distribution of test scores for a fixed individual is assumed to be binomial in form. The assumption of a binomial distribution is most tenable for dichotomously scored items, independent items, and items of equal difficulty. Huynh concluded that the model developed was "particularly suitable when testing is intermingled with instruction." (p. 263). Similar to a classical reliability index, k

increases as a function of test length and score variability. $\underline{k}$ also varies with cut-off score and has smaller values at both extremes (high and low) of the score range.

A similar model to that developed by Huynh (1976) is one proposed by Subkoviak (1976). The reliability estimate is called the coefficient of agreement and results from a single administration procedure for estimating the reliability of a criterion-referenced test. The procedure is based on two assumptions: (1) scores x_i and x'_i (scores on parallel tests) are independently distributed and (2) the distributions of x_i and x_i' for a fixed person are identically binomial in form. The coefficient of agreement, p_c, is an estimate of the extent to which students would be assigned to the same mastery states as a result of two test administrations.

Another recent approach to this issue was developed by Brennan and Kane (1977). They developed an index of dependability in the context of generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972). The index is quite similar to that developed by Livingston (1972) and consequently is based on a concept of the purpose of criterion-referenced (mastery) testing being to discriminate an examinee's universe score from a cut-off score. Brennan and Kane viewed the primary concern of mastery testing as whether an individual's universe score is above or below the criterion or cut-off score. As with Livingston's (1972) index, the index of dependability assumes a squared-error loss function.

In summary, the current state of the literature in criterion-referenced measurement and the reliability of criterion-referenced tests is extensive. Several models have been proposed in the last decade which fall into the following major categories:

Squared-error loss function: Livingston (1972), Brennan and Kane (1977).

Threshold loss function: Hambleton and Novick (1973); Swaminathan, Hambleton, Algina (1974).

Single administration: Huynh (1976), Subkoviak (1976).

The present research will investigate three reliability models: Swaminathan, Hambleton, Algina (1974), Huynh (1976) and Subkoviak (1976). A given data set will be analyzed to determine the robustness of each model to violations of underlying assumptions.

Chapter 3

STATEMENT OF THE PROBLEM

The Instructional System in Mathematics (ISM) is an objectives-based instructional management system for Grades K-8, in the Montgomery County (Md.) Public Schools (MCPS). ISM uses the MCPS computer to provide information for instructional management rather than for actual instruction. There are three major components to ISM: an instructional component, an assessment component, and a reporting component.

The instructional component is made up of a curriculum for Grades K-8 and instructional guides which accompany the curriculum. In the curriculum are approximately 190 key objectives which are organized by grade level. In addition, there are many other objectives which support the key objectives. The instructional guides identify the key objectives, provide assessment for these objectives and suggest instructional activities appropriate for the objectives.

The second part, the assessment component, is the one which is most important for the current research. The assessment component is made up of placement tests and mastery tests. Each type of test will be discussed separately.

There are three parts to the placement tests. There is a test for whole numbers concepts, a test for geometry and other related

in the target country. This last vehicle may be linked with the foreign state because of ideological affinity, mutual economic interests or ethnic identification. Political movements, trade unions, manufacturer associations, religious groups and ethnic groups all can play a role in middle-range penetration.

Karl Deutsch suggests in an article entitled "External Influences on the Internal Behavior of States" that the nation-state may be open to penetration and influence on some issues and not on others. The decision-making system may be highly centralized (and therefore less open to external influences) in regard to some operations than in regard to others. One subsystem of the national community may be linked to a foreign government and may be active on only particular issues. Deutsch defines a linkage group or potential linkage group (a subsystem in the national system) as "a group with links to the domestic system and with some particular links to the international or foreign input."[17] He suggests that this group is more likely to be susceptible to inputs from abroad if its ties to the domestic system are weakened, i.e., if it is discriminated against socially or economically, or if it perceives itself as such. The conceptual advantage of Deutsch's definition is that it lacks connotative implications. A linkage group may be loyal or disloyal to the nation-state, helpful or harmful to its interests.

The concepts of linkage and penetration were further discussed by James Rosenau and Wolfram Hanrieder.[18] Both limit the application of the concept of penetration to situations in which the penetrated nation-states are weak and vulnerable to outside "dominance" and the penetrator, relatively strong. Rosenau argues that "a penetrative process occurs when members of one polity serve as participants in the political processes of another: That is, they share with those in the penetrated polity the authority to allocate its values."[19] To illustrate, he provides the example of an occupying army, which is clearly a foreign body, making decisions and acting as part or as the only decision-making body of the occupied state. This is the most extreme case of penetration. He adds that the activities of foreign aid missions, subversive cadres on the

staffs of international organizations, the representatives of private corporations, the members of certain transnational political parties and others all can be linkages in a penetration process. His definition, however, unnecessarily restricts penetration activity to situations in which the participation is authoritative and direct. It is doubtful if most penetration activity could be considered "authoritative", following Robert Dahl's definition of authority as legitimate power or influence.[20] Certainly many citizens of the occupied polity and a substantial number of its leadership would not consider the foreign penetrator who has become the dominant force in the decision-making apparatus as "authoritative". Why do we need to insert the condition of "authority to allocate values" into the definition of the penetration process at all? Was the Vichy government perceived as legitimate by the majority of French citizens? Yet, was that not a very penetrated system? Who determines that a penetration is authoritative? Other members of a puppet government? the citizenry? the foreign penetrator? Do Arabs living on the Israeli occupied West Bank consider the Israeli government as "authoritative"?

Rosenau enumerates two other types of linkage processes, the reactive and the emulative. The reactive process, he states, is the reverse of the penetrative process. "It is brought into being by recurrent and similar boundary crossing reactions rather than by the sharing of authority. The actors who initiate the output do not participate in the allocative activities of those who experience the input, but the behavior of the latter is nevertheless a response to behavior undertaken by the former."[21] This type of linkage, according to Rosenau, is the most frequent and may be the result of direct or indirect activities. An example of this is the activities of Palestinian groups in Western Europe who have forged ties with parties of the extreme Left and with anti-government groups in an attempt to effect a change in government policy toward the Palestinian question. Another example was the impact of the rise of Nasser to the leadership of the pan-Arab nationalist movement in the late 1950s, early 1960s on the rise in anti-French demonstrations and revolution in Algeria in 1962.

The last type of linkage process is a special form of the reactive type. Rosenau calls it the emulative process which is essentially the "demonstration" effect whereby political activities in one country are perceived and emulated in another. Anti-war protests in the United States in the late 1960s were inspired by similar anti-government protest activities in several countries in Asia and Western Europe. Jerome Skolnick, who studied the student protest movement writes,

> The white student movement in America received inspiration in its early stages from dramatic student uprising in Japan, Turkey and South Korea ... American activists have been influenced by street tactics learned from Japanese students and by ideological expression emanating from France and West Germany ... The symbols of "alienated" youth culture, originating in Britain and the United States, have been adopted throughout Eastern and Western Europe ... The increasing cross-fertilization and mutual inspiration ... are then, the outcome of mass communication and informal contact.[22]

The reactive process corresponds roughly to what we have called "middle range penetration". The penetrator does not insert himself into the government apparatus but instead interacts with non-governmental actors with the intent of influencing or modifying their behavior (and thus influencing their government). Rosenau's attempt to delineate between this behavior and what he calls the penetrative process confuses more than clarifies. Both categories involve "penetration"; the former, of non-governmental actors, the latter, the formal decision-making apparatus. Both may influence decision-making either directly or indirectly and both, in my view, may be "non-authoritative". In fact, the penetrated state may regard the attempt to penetrate non-governmental actors as "legitimate" behavior and an attempt to establish a foothold within the government "illegitimate", i.e., non-authoritative![23]

Wolfram Hanrieder discusses the linkage between the external and internal dimensions of foreign policy. He criticizes the discipline for ignoring transnational phenomena and explains that the lack of linkage research in the past is because of "the tendency in the study of domestic politics to hold the

international environment constant, and a corresponding inclination in the study of international politics to hold the domestic environment constant."[24] He attempts in part to defend the separation between the "two levels of analysis"[25] by explaining the difficulties of correlating propositions derived from these two environments, since they stem from differently organized sets of empirical data and methodological assumptions.[26]

Hanrieder develops two concepts for foreign policy analysis -- compatibility and consensus. Compatibility attempts to measure "degrees of feasibility of various foreign policy goals, given the structures and opportunities of the international system." Consensus "assesses the measure of agreement on the ends and means of foreign policy on the domestic political scene."[27] He then tries to apply these two concepts in a linkage analytical framework utilizing the concept of a "penetrated system".

Hanrieder criticizes Rosenau's definition of penetration for its stress on "authoritative" participation. He says this limits the participation to institutions and ignores penetration by events or people that take place without being direct and authoritative. He suggests his own definition of a penetrated system: "A political system is penetrated a) if its decision-making process regarding the allocation of values or the mobilization of support on behalf of its goals is strongly affected by external events, and b) if it can command wide consensus among the relevant elements of the decision-making process in accommodating to these events."[28] He views penetration as the process through which the goals of the nation-state and its environment overlap. The definition describes the most extreme case of a totally penetrated system; however, empirically, penetration can range from totally penetrated to marginally penetrated. He suggests that it now becomes possible to analyze systems of linkage between the international system and national systems by "applying concepts that, although they originate from distinct analytical environments, are sufficiently isomorphic to allow cumulative propositions."[29] Thus, penetration can be measured on a

continuum -- and the degree of consensus in accommodating to the penetration will vary from situation to situation. Hanrieder's definition describes an "ideal type" of penetrated system and implies that other systems under varying circumstances may be more or less penetrated. Thus the clause that requires "wide consensus" among the relevant elements of the decision-making process refers only to a totally penetrated system. A less penetrated system may have sharp dissensus among the relevant elements of the decision-making process as a result of the penetrating factors from the environment. Partial penetration may be enough to weaken a domestic government's ability to make decisions, thus preventing it from taking any action.

The effects of the penetration may be felt only with respect to the limited issues which the penetrator actively pursued. One example of this is the penetration of Lebanon since 1971 by Palestinions who have in fact forced the Lebanese government into a state of submission on the question of border attacks into Israel and freedom of activity on its southern border. Yet they do not "control" the Lebanese government or dictate policy on issues of the economy, social welfare or most foreign relations. In the last few years, the influence of the Palestinians and their Shi'ite allies has increased, thus impeding the freedom of action of the Lebanese government in foreign policy. Further restricting the decision-making freedom of the Lebanese government is thepresence of another foreign element -- the Syrian peacekeeping force which has occupied part of Lebanon since 1979, ostensibly to maintain order and to separate between the Palestinians and the Christians. Lebanon is a penetrated system and is very much subject to the pressures and demands of the transnational "foreign" groups within it -- yet it could not be described as a totally penetrated system. The allocation of its values and goals are strongly influenced by the Palestinians and the Syrians, and its decision-making apparatus (for lack of power to do otherwise) has reluctantly agreed to accommodate to them.

A much less penetrated system is the United States, whose policies toward Israel and the Middle East are in part influenced by the activities of a linkage

(2) Huynh's $\hat{k}$ Coefficient

The second estimate which will be investigated is Huynh's (1976) estimate of kappa, $\hat{k}_H$.

$$\hat{k} = (\hat{p}_o - \hat{p}_c) / (1 - \hat{p}_c)$$

where $\hat{p}_o$ = estimated proportion of consistent classifications on two test administrations using Huynh's procedure.

and

$\hat{p}_c$ = estimated expected proportion of consistent classifications, on two test administrations, again using Huynh's procedure.

Huynh (1976) developed this model for estimating reliability when data are only available for a single test administration. Thus, one must estimate p_{oH} and p_{cH}.

In order to explain Huynh's (1976) estimate it is necessary to examine the model which he proposes and the assumptions of the model. Huynh (1976) develops the reliability estimate within the framework of the beta binomial model. The applicability of the beta binomial model to criterion-referenced testing situations has been investigated by Gross and Shulman (1978).

There are two basic assumptions which must be met when using the beta binomial model. The examinee's score or number right on the criterion-referenced test (x) is used to infer the true ability of the examinee (θ). The first assumption is that the ability parameter (θ) is distributed within the examinee population as a beta random variable with parameters α and β .

The assumption of a beta distribution for the ability parameter appears reasonable since the family of beta distributions can assume many different shapes - e.g., normal, skewed etc. When data are available for a large number of examinees, both α and β can be estimated from the test score distribution. If μ and σ are the mean and standard deviation of the test score distribution and if the KR21 reliability is denoted by

$$KR21 = \left(\frac{n}{n-1}\right)\left(1 - \frac{(n-\mu)}{n\,\sigma^2}\right) \quad \text{where } n = \text{number of items on the test} \qquad (1)$$

then

$$\alpha = (-1 + {}^{1}/_{\alpha_{21}})\,\mu \qquad \text{and} \qquad (2)$$

$$\beta = -\alpha + {}^{n}/_{\alpha_{21}} - n \qquad (3)$$

The second assumption is that the conditional distribution of $\underline{x}$ (examinees's test score) is a binomial distribution with parameters $\underline{n}$ and θ. The necessary conditions for this to be a reasonable assumption are: (1) each test item is scored 0 or 1; (2) the test items in the universe are exchangeable - e.g., the test score distribution based on n items does not depend on which items happen to be chosen, implying that the items are statistically independent so that outcome on one does not affect outcome on the others and thus, the probability of a correct response remains constant across items. While the assumption of similar item difficulty may not be tenable in a criterion- referenced testing situation, Gross and Shulman (1978) and Subkoviak (1978) report that the beta binomial model appears to be robust with respect to violations of equal p values for the items. The issue of violation of item homogeneity will be addressed in the present research by

constructing tests which are the same length but which contain items measuring different level objectives i.e., objectives for different math skills. This process is described in detail in an upcoming section.

Within the beta-binomial framework the distribution of test score x is beta binomial in form

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\ \Gamma(\beta)} \binom{n}{x} \frac{\Gamma(\alpha + x)\ \Gamma(\beta + N - x)}{\Gamma(\alpha + \beta + n)} \qquad (4)$$

where x = 0, 1, 2 . . . n, and $\Gamma(\cdot)$ is the gamme function.

When dealing with equivalent forms of a test, Form X and Form Y, Huynh (1976) shows that X and Y follow a bivariate beta binomial distribution with joint density

$$f(x,y) = \frac{\binom{n}{x}\binom{n}{y}}{B(\alpha,\beta)}\ B(\alpha + x + y,\ 2n + \beta - x - y), \qquad (5)$$

where $B(2,\beta)$ is the beta function.

In focusing on the binary results of achieving mastery if x is at least c (cut-off or criterion score) and not achieving mastery otherwise, Huynh (1976) denotes the proportion of examinees achieving mastery on both forms as p_{11}, given by

$$p_{11} = \sum_{x,y=c}^{n} f(x, y) \qquad (6)$$

and the proportion of examinees classified as masters on one form only as

$$p_1 = \sum_{x=c}^{n} f(x) \qquad (7)$$

Given that and have been estimated using equations (2) and (3) one can construct the joint distribution for x and y,

$f(x,y)$. Estimates of p_o, p_{11} and thus k can then be obtained. The entire procedure can be outlined as follows:

1. Compute the mean ($\hat{\mu}$), variance ($\hat{\sigma}^2$), and Kuder-Richardson coefficient 21 ($\hat{\alpha}_{21}$) of the scores on Form X.

2. Compute parameters $\hat{\alpha}$ and $\hat{\beta}$ which, together with the number of test items, determine the shape of the joint distribution of scores on the two forms $\hat{\alpha} = -1 + \frac{1}{\hat{\alpha}_{21}} \mu$ and $\hat{\beta} = -\alpha + \frac{1}{\hat{\alpha}_{21}} - n$

3. Using the values of $\hat{\alpha}$, $\hat{\beta}$ and n determine the joint distribution of scores on Forms X and Y. This distribution is symbolized $\underline{f}\ (\underline{x}, \underline{y})$ which represents the probability of persons scoring $\underline{x}$ on Form X and $\underline{y}$ on Form Y. Given the values of $\hat{\alpha}$ and $\hat{\beta}$ for a test of a given length the value of $\underline{f}\ (\underline{x}, \underline{y})$ for scores $\underline{x} = 0$ and $\underline{y} = 0$ can be obtained as follows:

$$f(0,0) = \prod_{i=1}^{2n} \left[(2n + \hat{\beta} - i) \,/\, (2n + \hat{\alpha} + \hat{\beta} - i)\right] \qquad (8)$$

For example, if $\hat{\alpha} = 12.52$, $\hat{\beta} = 14.52$ and $n = 10$, (Subkoviak, 1978) the value of $f(0, 0)$ is:

$$f(0,0) = \prod_{i=1}^{2n} \left[(2n + \hat{\beta} - 1)/(2n + \hat{\alpha} + \beta - i)\right]$$

$$= \left[\frac{20 + 14.52 - 1}{20 + 12.52 + 14.52 - 1}\right] \cdot \left[\frac{20 + 14.52 - 2}{20 + 12.52 + 14.52 - 2}\right] \cdot \cdot \cdot$$

$$\cdot \frac{20 + 14.52 - 20}{20 + 12.52 + 14.52 - 20}$$

$$= .002$$

Then, given $\hat{f}(0, 0) = .002$, values of $f(x, y)$ for other x and y pairs, are obtained as follows:

$$\hat{f}(x + 1, y) = \hat{f}(x, y) \cdot \frac{(n - x)(\hat{\alpha} + x + y)}{(x + 1)(2n + \hat{\beta} - x - y - 1)} \quad (9)$$

It should be noted that $\hat{f}(x, y)$ is symmetric in the sense that $\hat{f}(x, y)=\hat{f}(y, x)$. After computing, the joint distribution of scores on the two forms can be obtained and the proportion of examinees that would obtain score x and Form X and score y on Form Y can be entered into a table (see Table 23 in the Appendix).

To continue this example, it can be shown how to estimate $\hat{k}$ using the table of the joint distribution of x and y.

4. Coefficient $\hat{p}_o$, the proportion of consistent classifications is obtained by summing appropriate entries in the joint distribution table. Given a particular cut-off score the proportion of persons consistently classified as masters is obtained by summing the probability values for all pairs which are at or above the cut-off score. Similarly the proportion of persons consistently classified as nonmasters on both tests can be obtained by summing the probability values for all pairs below the cut-off score. Then, the total proportion of consistent decisions is the sum of the proportion of persons consistently classified as masters and the proportion of those consistently classified as nonmasters.

5. Coefficient $\hat{k}_H = (\hat{p}_{oH} - \hat{p}_{cH})/(1 - \hat{p}_{cH})$ which is the proportion of consistent decisions beyond that expected by chance can also be obtained from the joint distribution table.

Proportion p_{oH} was obtained in the previous step. p_{cH}, the proportion of consistent decisions due to chance, is a function of the marginal proportions of masters and nonmasters in the joint distribution table. For a given cut-off score, the proportion of masters expected by chance is the sum of all the columns of probability values for the cut-off socre and above. The proportion of nonmasters is 1 - proportion of masters expected by chance. Therefore $\hat{p}_{cH} = \Sigma\, \hat{p}_{k.}\ \hat{p}_{.k}$.

6. Finally, the proportion of consistent decisions beyond that expected by chance is

$$\hat{k}_H = \frac{\hat{p}_{oH} - \hat{p}_{cH}}{1 - \hat{p}_{cH}} \tag{10}$$

(3) Subkoviak's Coefficient of Agreement

The third reliability estimate which will be examined is that developed by Subkoviak (1976), also intended for single test administration data. Subkoviak (1976) calls his estimate a coefficient of agreement and the coefficient of agreement $p_{c(s)}$ for a group of n persons when the criterion equals c is written

$$\hat{p}_c(s) = \frac{\sum_{i=1}^{N} \hat{p}_{cs}^{(i)}(s)}{N} \tag{11}$$

The coefficient of agreement for a group of N persons is defined as the mean of the individual coefficients of agreement $p_{cs}^{(i)}$, where $p_{cs}^{(i)}$ is the probability that individual i is assigned to the same mastery state on parallel tests x_i and x'_i.

$$p_{cs}^{(i)} = p(x_i \geq c,\ x'_i \geq c) + p(x_i < c,\ x'_i < c) \tag{12}$$

There are two assumptions which make possible the estimation of the individual coefficient and thus the estimation of the group coefficient. The first assumption is that the scores x_i and x'_i are independently distributed for a fixed person i. The implication of this assumption is that the experience of taking test x does not affect the outcome on test x' for person i. The second assumption is that the distributions of x_i and x'_i for a fixed person are identically binomial in form. As was described with Huynh's (1976) model the necessary conditions for this to be a reasonable assumption are (1) each test item is scored 0 or 1; (2) the outcome on one item does not affect the outcome on another and (3) the probability of a correct response remains constant across items.

Under the assumption that the distributions of x_i and x'_i for a fixed person are identically and independently binomial in form, the preceding equation for $p_{cs}^{(i)}$ simplifies to

$$p_{cs}^{(i)} = p(x_i \geq c)^2 + 1 - p(x_i \geq c)^2 \qquad (13)$$

where

$$p(x_i \geq c) = \sum_{x_i=c}^{n} \binom{n}{x_i} \hat{p}_i^{x_i} (1 - \hat{p}_i)^{n - x_i} \qquad (14)$$

As with Huynh's (1976) model, it can be assumed that the binomial model is robust with respect to violations of item homogeneity which occur on actual tests. The quantity $\hat{p}_i$ is the estimated true probability of a correct item response for person i, which can be estimated from his/her observed score

x_i on a single test i.e., $\hat{p}_i = {}^{x}i/n$, where x_i is number correct and n is the total number of items. Subkoviak (1976) points out that this should lead to reasonably accurate results if $n > 40$. Since in the current research and indeed in much locally developed criterion-referenced testing programs $n < 40$, it is important to demonstrate a more appropriate estimate of p_i for small n.

A better estimate of p_i for small n is given by:

$$p_i = \alpha_{21/x}\left(\frac{x_i}{n} + (1 - \alpha_{21/x})\left(\frac{M_x}{n}\right)\right) \text{ where} \tag{15}$$

$$\alpha_{21/x} = \frac{n}{n-1}\left[1 - \frac{M_x\,(n - M_x)}{n\,S_x^2}\right] \tag{16}$$

and where, $\alpha_{21/x}$ is the Kuder-Richardson Formula 21 reliability coefficient, M_x is the mean of the test score distribution, and s_x^2 is the variance of the distribution.

Summary

In summary, Swaminathan et al.'s (1974) kappa can be used as a standard to which to compare the estimates of Huynh (1976) and Subkoviak (1976). When data are available from two test administrations, kappa can be calculated empirically without any restrictive assumptions. On the other hand, the two reliability estimates based on a single test administration are valid only to the extent that the underlying assumptions of the models are met. The Huynh (1976) and Subkoviak (1976) estimates have some similarities in their assumptions as well as in their simulations of scores on the second form of the test. Both use the KR21 in their simulations, both use the mean of the test score distribution and the difference between the mean and the number of items on the test.

The primary difference between the two estimates is that there are more distributional assumptions for Huynh's (1976) estimate. Huynh's (1976) model assumes a binomial distribution of the number right given ability and further that the ability parameter has a beta distribution. Subkoviak's (1976) model may be more robust as it assumes only a binomial distribution of number right, given ability. Since the beta distribution can take on so many forms, however, this distinction between the models may not be particularly significant.

PREDICTIONS OF BEHAVIOR OF RELIABILITY ESTIMATES

The value and/or standard error of a traditional or norm-referenced reliability estimate varies as a result of manipulating test length, sample size, ability of the examinees, and heterogeneity of test content. Part of the current research will investigate how these factors affect reliability estimates for criterion-referenced tests. In addition to the aforementioned factors, criterion-referenced reliability estimates might vary as a result of manipulating the criterion score. The following section will describe in detail the manipulations of the data in order to investigate the effect on the reliability coefficients. However, this section will propose some hypotheses about the behavior of the criterion-referenced coefficients, under different conditions. The predictions about each coefficient will be grouped together under the condition which is varying.

I. Test Length

Since kappa is concerned with consistency of decision-making rather than deviation of scores from the criterion score, it is not totally clear what the effect of test length on kappa would be. Since n, test length or number of test items, does not directly enter into the computation of kappa, its effect is not clear. However, a longer test is probably a more reliable test. Logsdon (1979) found that kappa values revealed no pattern of relationship to test length, in a study involving the administration of alternate forms of a criterion-referenced reading comprehension test in a rural school district. Using four criterion levels for tests of 12, 24, 36 and 48 items,

Logsdon (1979) found that mean kappa values were generally higher for the 48-item tests than for the 12-item tests but that the mean values for the 24- and 36-item tests were extremely variable and did not follow a general pattern of increase. However, Huynh (1978) reported in a theoretical paper on the reliability of multiple classifications, that kappa increases with test length within both the normal test score model and within the beta-binomial model. In Logsdon's (1979) study he did report large standard errors for kappa ($\sim$.07 to .16) and thus the standard errors may be obscuring the relationship between k and increasing test length which Huynh (1978) reported.

$\hat{k}_H$ (Huynh's estimate of k) is conceptualized as a single sample estimate of k. If $\hat{k}_H$ is a "good" estimate of k, its behavior with respect to test length should be similar to the behavior of k itself. Thus $\hat{k}_H$ is predicted to increase as a function of test length, and to provide a good estimate of k at all test lengths.

With respect to Subkoviak's coefficient of agreement ($p_{c(s)}$), since the binomial distribution assumption underlies this coefficient as it does $\hat{k}_H$, it seems likely that p_{cs} would increase with test length. However, it is not clear what the exact effect of increasing test length will be on this coefficient.

II. The Combined Effect of Cut-Off Score and Test Length

It is unclear how changing the cut-off score for different test lengths will affect k.

In a comparison of three data sets representing different degrees of homogeneity of test scores, Huynh (1976) reported that $\hat{k}_H$ starts increasing as the cut-off score takes larger values, up to a point.

When $\hat{k}_H$ reaches a maximum, it then decreases. Huynh (1976) explains this by noting that p_c is close to 1 when the cut-off score is too small or too large, meaning that there is not much opportunity for improving the consistency of decisions beyond the chance level. If this is the case with these data, one would expect $\hat{k}_H$ to reach its highest values for a cut-off score of 80 percent and decline for the 100 percent cut-off score.

It is also unclear how changing the cut-off score for different test lengths will affect $p_{c(s)}$. Since the method involved in computing $p_{c(s)}$ involves the assumption of a binomial distribution of observed scores, it is possible that the behavior of $p_{c(s)}$ under these conditions will be similar to that of $\hat{k}_H$.

III. The Combined Effect of Ability Grouping and Cut-Off Score

For different levels of student ability, kappa should change as a result of changing the mastery level. The relationship which k has to cut-off score depends on ability. Thus, it is predicted that for extreme ability groups, high or low, the change in cut-off score will not have much effect on k, but for a heterogeneous ability group it will. The reason for this is that for very high ability students one would expect consistent classification as masters regardless of cut-off score and similarly for low ability students one would expect consistent classification as nonmasters regardless of cut-off score.

With respect to $\hat{k}_H$, heterogeneous ability casts doubt on the beta distribution assumption. Therefore, for homogeneous ability groups, such as high or low ability, $\hat{k}_H$ would be expected to provide a good estimate of k.

Similarly for $p_{c(s)}$ heterogeneous ability may violate the underlying assumptions and so for a homogeneous ability group such as high or low ability, $p_{c(s)}$ would be expected to provide a different estimate of reliability than for a heterogeneous group.

IV. Sample Size

There is no effect from sample size which can be predicted for kappa. The two sample sizes to be examined are one classroom ($N \approx 25$) and two classrooms ($N \approx 50$). The size of the sample will probably not affect this formula at all since N does not enter directly into the computation of k. The standard error, however, should depend on N, i.e., it should decrease as a function of increasing N.

With respect to $\hat{k}_H$, the only effect from sample size which can be predicted is on the standard error. The same prediction holds for $p_{c(s)}$.

In terms of the "goodness" of the estimates of reliability, the size of the sample should have no effect on the accuracy of estimation of k. Therefore, the prediction about the accuracy of the estimates of k is the same as the prediction for the combined effect of ability grouping and cut-off score - e.g., $\hat{k}_H$ will not be good a estimate of k when the ability of the students is heterogeneous, regardless of sample size.

V. Heterogeneity of Test Content

It is not clear what effect this factor will have on kappa. The reason for including this variable is that it leads to violations in the distribution assumptions underlying the Huynh (1976) and Subkoviak

(1976) estimates. Heterogeneous test content makes the binomial distribution assumption for the number correct untenable. Thus when we have heterogeneous ability (which casts doubt on the beta distribution assumption of Huynh's (1976) model and which was the case in this analysis) and heterogeneous item content one may hypothesize that $\hat{k}_H$ will provide a poor estimate of k.

VI. Comparison of Norm-Referenced Reliability Coefficients (KR21 and Test-Retest) with the Criterion-Referenced Reliability Coefficients

It is unclear what the relationship between the Kuder-Richardson coefficients and k, $\hat{k}_H$ and $p_{c(s)}$ will be. The same factors of test length, cut-off score, sample size, examinee ability and heterogeneity of test content may affect the KR21 coefficients in the same way as those factors affect the criterion-referenced coefficients. However, since the KR21 coefficients assess the internal consistency of tests while the criterion-referenced coefficients assess decision consistency in a test-retest situation, there may be little relationship between these two types of coefficients.

Since k is essentially a test-retest coefficient and $\hat{k}_H$ is an estimate of this test-retest coefficient one may hypothesize a strong relationship between k, $\hat{k}_H$ and the test-retest coefficients. One may also hypothesize a strong relationship between $p_{c(s)}$ and the test-retest coefficients since $p_{c(s)}$ assesses consistency of mastery decisions.

Before summarizing the hypotheses about the behavior of the reliability coefficients, it should be noted that the scaling of the three criterion-referenced reliability estimates is different. k and

$\hat{k}_H$ are scaled in the same way while the scaling for $p_{c(s)}$ is different. The coefficient $p_{c(s)}$ does not correct for chance agreement i.e., agreement due to the proportion of masters and nonmasters in the group and thus is indicative of the total proportion of consistent mastery classifications which occur on two tests. Coefficients k and $\hat{k}_H$ on the other hand take out the part of the consistency which is due to the mastery/nonmastery composition of the group taking the test (p_c). Coefficients k and $\hat{k}_H$ thus can be considered measures of test consistency while $p_{c(s)}$ is a measure of test consistency as well as consistency resulting from the proportion of masters and nonmasters in the group tested. Given this fact the size of the $p_{c(s)}$ coefficient will always (theoretically) be larger than the k or $\hat{k}_H$ coefficients.

This fact can be explained as follows. Two components of the overall consistency as measured by p_o are the accuracy of the test and the numbers of masters and nonmasters in the group tested. If for example using a reliable test the group being tested consists primarily of high ability students, the proportion of consistent mastery/mastery outcomes would be high. Similarly a group composed primarily of low ability students would result in consistent nonmastery/nonmastery outcomes. In either situation, the number of consistent decisions would be high and the value of $p_{c(s)}$ would also be high. However, the use of $p_{c(s)}$ would not enable one to separate what portion of the consistency was attributable to the group being tested and what portion was attributable to the test.

Coefficient k extracts from p_o the consistency which is attributable to the numbers of masters and nonmasters in the group. In

the formula for $k = (p_o - p_c)/(1 - p_c)$, p_c is the portion of consistency which is due to the numbers of masters and nonmasters in the group tested. The calculation of k removes p_c from the total proportion of consistency p_o. The resulting value is the consistency of the test. Since $\hat{k}_H$ is an estimate of kappa based on one test administration, the portion p_c is also removed from the total consistency p_o, and $\hat{k}_H$ is an estimate of the consistency of the test.

In summary, the following hypotheses about the reliability estimates are formulated:

H_1: All three reliability estimates, k, $\hat{k}_H$ and $p_{c(s)}$ will increase as the tests are lengthened.

H_2: $\hat{k}_H$ will provide a good estimate of k at all test lengths.

H_3: For groups of high ability students or low ability students, the change in cut-off score will have little effect on any of the three coefficients, k, $\hat{k}_H$ or $p_{c(s)}$.

H_4: $\hat{k}_H$ should provide a good estimate of k for homogeneous ability groups.

H_5: The size of the sample will affect only the standard errors of all three reliability coefficients.

H_6: For heterogeneous test content both $\hat{k}_H$ and $p_{c(s)}$ will be lower than for homogeneous test content.

H_7: $\hat{k}_H$ will provide a poor estimate of k for heterogeneous test content.

H_8: The KR21 coefficients will have little relationship to any of the three criterion-referenced reliability estimates.

H_9: The test-retest coefficients will show a moderate to strong relationship to all three criterion-referenced reliability coefficients.

Thus, it is anticipated that the present research will result in some guidelines for the use of the three criterion-referenced coefficients. Since kappa represents an ideal standard which one would want to compute whenever there could be two test administrations, the comparison of the two estimates $\hat{k}_H$ and $p_{c(s)}$ to kappa should give an indication of each estimate's utility. In cases where there can only be one test administration the examination of the effect on the estimates of the variables of test length, cut-off score, ability level of the students, sample size and heterogeneous test content should reveal the sensitivities of each. Thus one estimate would be able to judge the utility of each index for a particular testing situation.

Finally, the comparison of the criterion-referenced reliability estimates with norm-referenced coefficients should provide some insight into whether norm-referenced coefficients provide any information about criterion-referenced test data. This comparison also may indicate differences in the information provided by the two types of indices. These pieces of information should be beneficial to educational practitioners for the development and validation of criterion-referenced instruments.

Chapter 4

METHODOLOGY

In order to investigate the properties and the usefulness of the three reliability estimates under different conditions, it is necessary to have test data from two test administrations. Therefore, data have been collected in the following way:

(1) A random sample of classes in Grades 3, 5, 6, 7, 8 in the ISM program was chosen for test administration. The total sample size is 325 students with approximately equal numbers at all five grade levels. The number of students taking each objective differ, however. See Table 20 in the Appendix) for the number of students taking each objective.

(2) Students in each classroom took the placement test in Whole Numbers--W-1 for Grade 3 and W-2 for Grades 5-8 as well as two mastery tests, one test measuring one multiplication objective and one test measuring one division objective (The 3rd grade classes took only one mastery test in multiplication). Each mastery test administered to the sample classes assesses an objective which is also assessed on the placement test. Therefore, each student in the sample took one placement test (W-1 or W-2) consisting of between 60 and 70 items and two mastery tests, each one consisting of five items and each one measuring an

objective which is also measured on the placement test. The one exception to this is the Grade 3 sample as the multiplication mastery test (MU05-H) has 10 items.

(3) In order to control for the effects of intervening instruction, the data collection dates were determined by when students had a scheduled break in instruction. The tests were administered during the week prior to the spring vacation and during the week following the spring vacation resulting in an interval of two weeks between the two test administrations. All students who took the tests before spring break (each student took one placement test and two mastery tests at each sitting) took the same tests in the same order during the second test administration. The testing occurred during the spring of 1979.

The rationale for the data collection as outlined above is based on several facts.

The Whole Numbers Placement Test (W-2) was chosen because it is intended for a wider range of students (4 grade levels) than the other placement tests, and it should allow for finding an adequate number of students who are more proficient and an adequate number of students who are less proficient in the skills tested. In the Montgomery County Public Schools the students who are the highest achievers in math take either Algebra 1 or Unified Math (advanced math) when they enter Grade 7. Therefore, students who are in ISM in Grades 7 and 8 are often students who are less proficient in Math. The items which measure MU 05-H on the placement test W-1 are

the same items which are on the placement test W-2. Therefore, the 3rd grade classes which took W-1 and also took the MU 05-H mastery test are included in this analysis. This will add to the analysis because it will increase the number of objectives which can be included and it will add an interesting perspective on performance of different grade level students on these items when they are embedded in different tests. On the W-1 test, these are advanced items while on the W-2 test, they are easier items.

Secondly, mastery tests were chosen so that they each measured one of the same objectives which appeared on the placement test. Since the test contained a large number of multiplication and division objectives, one mastery test was selected for a multiplication objective, and one was chosen for a division objective. This selection was made so that the effects on the reliability estimates of pooling items from the several objectives could be investigated. Pooling, however, will be done only <u>within</u> category, either multiplication <u>or</u> division, but not across the two categories.

Tables 20 through 22 (in the Appendix) present the data on numbers of items on the various tests administered and the numbers of students taking each test.

The W-1 Test is a Placement Test in Whole Numbers, designed for students in Grades 3 and 4. This test contains items which assess objectives in Numeration, Place Value, Addition, Subtraction, Multiplication, and Division. However, the only items which are of

interest in this research are those which assess the multiplication objective MU 05-H. See Table 21 for the MU 05-H items from the W-1 Placement Test and from the MU 05-H mastery test.

The W-2 Test is a Placement Test in Whole Numbers designed for students in Grades 5-8. This test contains items which assess the same categories of objectives as those of W-1, but the levels of the objectives are more advanced. Only the multiplication and division items are of interest to this research. See Tables 22 and 23 for the multiplication and division items from the W-2 Test.

Tests which measure only one objective are called mastery tests and each one contains five items. The following tests are mastery tests: MU 05-H, MU 07-K, MU 08-L, DI 08-J, DI 10-N. See Appendix for copies of these tests.

Analyses

There are eight analyses which were undertaken using the ISM data collected in April 1979 and discussed at the beginning of the chapter (see Tables 20-23). The three reliability estimates were examined under different data manipulations to see whether they behaved as one would predict and then, the estimate of k, $\hat{k}$, (Huynh (1976)) was compared to the Swaminathan, Hambleton, Algina (1974) computation of kappa, under all data manipulations.

The reason for comparing the estimate to the computed kappa is that kappa is truly an empirical measure. One needs to have data from two test administrations in order to compute kappa and in this research there were two test administrations. Thus kappa is an

ideal standard. The estimates from a single test administration are valid only to the extent that the underlying assumptions of the model are met. Therefore, one can compare it to kappa when the assumptions have been met and when the assumptions have been violated.

The three reliability coefficients were each calculated using individual Fortran programs. The kappa coefficient (Swaminathan, Hambleton, Algina (1974) was computed by means of a program written specifically for this research.

The estimate of kappa (Huynh, 1976) was computed using a program written by Huynh in December 1979.

The coefficient of agreement (Subkoviak, 1976) was computed using a program written by Subkoviak and Albrecht (1979).

The computer program to compute the coefficient of agreement ($p_{(cs)}$) provides two options for computing the coefficient - one is based on the assumption that the distribution of an individual's scores across repeated testings is simple binomial and the other assumes the distribution is compound binomial. Since the assumption of the simple binomial model (see Chapter 3) may not be tenable, Subkoviak (1976) prefers the compound binomial model. He states that the actual score distribution for a person would probably be less variable than the simple binomial resulting in an assignment consistency which exceeds that predicted by the simple binomial model. The compound binomial model seems to allow the probability of success to vary across items and thus the compound binomial model is preferred. For this research, however, since the opportunity was

available to compute the p_{cs} coefficient based on either model, both models were specified to allow comparison of the results of the two models. Thus all discussions of results will include the results from both the simple and the compound models.

Each data manipulation and the subsequent reliability determinations and comparisons will be discussed in separate subsections.

A. Test Length

In order to investigate the effect of test length on the three reliability estimates, the data (test items) from one objective (such as MU 05-H, MU 07-K or MU 08-L) - where there are items from the placement test and the mastery test - were reorganized to construct tests of different lengths. For example, for the multiplication objective MU 07-K, there are five items on the placement test and five items on the mastery test of the same objective. Thus for the 100 students who took the items measuring MU 07-K on both the placement test and the mastery test, there are a total of ten items which can be manipulated into tests of different lengths.

While all the items which measure one particular objective are supposed to be appropriate measures of the objective, it is possible that the items vary in difficulty. In order to control for this potential variation, all possible combinations of items for a test of a given length were constructed from the pool of items measuring one objective. To continue the example of objective MU 07-K mentioned above, in order to look at a 5-item test from the 10 items

measuring this objective, a program was written to construct all possible combinations of a specified number of items. If n = total number of items and p = chosen test length; the total number of combinations (Comb) is given as:

$$\text{Comb} = \binom{n}{p} = \frac{n!}{p!\,(n-p)!} \qquad (17)$$

In the present study the number of combinations ranged from 1 to 1,763.

Thus, for each objective analyzed - MU 05-H, MU 07-K, MU 08-L, DI 08-J, and DI 10-N - all possible combinations of 5 items and 7 items (and where possible 10 items) were constructed. These combinations were constructed for the pre-tests and the post-tests.

Each analysis was conducted by grade level since different grades took different objectives. Across grade levels the test lengths examined differed according to the total number of items available for a particular objective. The test lengths examined by grade level were:

Grade 3 - multiplication: 5 items, 7 items, 10 items, 13 items

Grade 5 - multiplication: 5 items, 7 items, 10 items

Grade 5 - division: 5 items, 7 items, 8 items

Grades 6-8 - multiplication: 5 items, 7 items, 9 items

Grades 6-8 - division: 5 items, 7 items, 8 items

For each test length for each objective - (3 multiplication objectives and 2 division objectives) - all three reliability coefficients were computed.

In the analysis of the effect of test length on the reliability coefficients, mastery level had to be held constant. Since the numbers of items to be considered for different test lengths could

not always be multiplied by the same percentage and result in a whole number of items, the decision was made to choose the number of items for mastery which most closely matched the mastery level selected. The test developers set 80 percent as the mastery level for most of these tests. Therefore, the numbers of items selected for mastery level are those which are closest to 80 percent and as can be seen from the following table, all mastery levels are within 5 percent of the 80 percent level chosen by the test developers.

Total Number of Items on Tests	Number Items for Mastery	Percentage of Total e.g., Mastery Level
5	4	80
7	6	85
8	6	75
9	7	77
10	8	80
13	11	84

Swaminathan, Hambleton, and Algina's (1974) kappa was computed using the pre-test and post-test data. The Huynh (1976) and Subkoviak (1976) estimates were computed on the pre-test data. All three reliability estimates were examined in light of the predictions made about how they would behave as a result of test length manipulation. Then, the estimate of k was compared to the standard, kappa. In all cases but one (13 items) where there was more than one test of a particular length the mean coefficient and standard error were computed.

B. Cut-off Score

In order to investigate the combined effect of cut-off score and test length, the three reliability coefficients were computed for tests of different length using different mastery criteria. For each test length (5-item, 7-item, 8-item, 9-item, 10-item, 13-item), the mastery levels which were examined were 60 percent, 75 percent or 80 percent, and 100 percent. What this means is that all three reliability coefficients were computed for each mastery level and each test length. For example, for each 5-item test, the three reliability coefficients were calculated three times, once for each mastery level. In cases where there was more than one test of a certain test length, the mean estimate and standard error were computed for each mastery level. In cases where there was only one test of a certain length (e.g., 13 items), each reliability coefficient is reported for each mastery level (see Table 10).

As was the case with holding mastery level constant, the fact that there were several test lengths for which one cannot determine all the exact mastery levels chosen for investigation posed some difficulty. Thus, where possible, mastery criteria were chosen such that an exact number of items corresponded to an exact mastery level. In cases where this was not possible (7 items, 9 items, 13 items) mastery levels were chosen to be as close as possible to the

prespecified mastery levels. For the mastery level analyzed, see the following table:

Number of Items in Test	Prespecified Mastery Level	Number of Items and Percentage Mastery
5	60%	3(60%)
	80%*	4(80%)
	100%	5(100%)
7	60%	4(57%)
	80%*	6(85%)
	100%	7(100%)
8	60%	5(63%)
	80%*	6(75%)
	100%	8(100%)
9	60%	6(67%)
	80%*	7(78%)
	100%	9(100%)
10	60%	6(60%)
	80%*	8(80%)
	100%	10(100%)
13	60%	8(61%)
	80%*	11(84%)
	100%	13(100%)

Each reliability estimate was examined to try to determine if changing the cut-off score has the effect predicted in the previous section. Secondly, the estimate of k was compared to the standard, kappa, to see how it behaved relative to kappa at each cut-off score.

---

*The cut-off score of 80 percent was examined for the test length analysis.

C. Ability Level

In order to investigate the effect of ability of the test taker on the reliability estimates, scores from the Iowa Test of Basic Skills (ITBS) were used to identify students of high ability and low ability in mathematics. The analyses of test length and mastery criteria were done again within the low ability and the high ability groups. In each objective category a test of 5 items, 7 items and then the maximum number of items for that objective were selected for an examination using the low ability group and the high ability group. The maximum number of items for the objectives ranged from 8 items (for each of the two division objectives) to 13 items (for one multiplication objective). For the test length of 5 items and 7 items, the combination program was run so that all possible combinations of 5 items and the 7 items were evaluated. The mastery criteria examined were 60 percent, 80 percent, and 100 percent. Since there were thousands of 5-item and 7-item tests for each objective, the mean estimates and standard errors were computed.

D. Sample Size

Another variable which may affect the reliability estimates is the size of the sample. In order to investigate the effect of sample size on the reliability estimates, the analyses for test lengths and cut-off score were run again for different size samples. That is, the different length tests and the different mastery criteria were analyzed for sample sizes of 1 classroom ($\sim$ 25 students) and then the maximum number of students on that grade level which took the items for a particular objective. For the third grade and fifth grade

classes the maximum number of students which took the tests was 2 classrooms ($\sim$ 50 students). There were 200 students in Grades 6-8 who took the two objectives. For consistency with the analyses for Grades 3 and 5, Grade 7 was selected for an analysis of one and two classrooms. The Grade 7 data were then compared to the data for the total group of Grades 6-8. For the analyses undertaken, see the following table:

Objective	Grade	Number of Classrooms	Number of Students
MU 05-H	3	2	52
MU 07-K	5	2	56
DI 08-J	5	2	56
MU 08-L	7	2	58
DI 10-N	7	2	58

As with the analysis for ability level, in each objective category, a test of 5 items, a test of 7 items and then a test using the maximum number of items for that objective were selected for examination using the different sample sizes. The maximum number of items for the objective ranged from 8 (for the two division objectives) to 13 (for one multiplication objective). For the test lengths of 5 items and 7 items the combination program was run so that all possible combinations of 5 items and 7 items were evaluated. For these tests, the mean estimates and standard errors were calculated.

E. Item Heterogeneity

An examination of the effect of item heterogeneity was undertaken by controlling for test length and manipulating the types of items on the tests. This was accomplished by constructing

tests of the same length using items from different levels of objectives. For example for the 200 students who took the test of Multiplicaton 08-L, there are data about their performance on a test of 5 items. Therefore, for each of these students one can look at their performance on a composite test which consists of one item from each of the objective levels MU 04-H, MU 05-H, MU 06-I, MU 07-K and MU 08-L. These students have taken several items from each of these objectives by taking the placement test. Thus, one can examine the effect of mixing item types without confounding the procedure with test length varying also. Another example might be the 50 students who took the 10-item mastery test of MU 05-H. Data are available for their performance on a 10-item test at this level. However, how would they perform on a 10-item test which contained items from several objectives? One can determine this by constructing a 10-item test for these students using items from level MU 05-H, items from level MU 04-H and items from level MU 06-I (data which come from the placement test W-1 which these students took). The issue of test heterogeneity can be examined in this way which is a procedure of redefining over and over what one means by a test.

In order to conrol for the effect of variation due to particular heterogeneous items selected, a computer program was written to select combinations of items for a specified test length from among the items at the different objective levels. Since all possible combinations of items for a specified test length often meant millions of combinations, a program for a random number generator

was included in the program for selecting items. Thus for an examination of item heterogeneity, each analysis consisted of 100 possible combinations, randomly selected. For each objective, then, tests of the same length as those investigated in the test length analysis, were constructed using heterogeneous items. This means that for each objective, 100 combinations of 5 items, 7 items and where possible 10 items or 13 items were constructed using items from several different objective levels.

The total number of items available for constructing into heterogeneous tests were different at the different grade levels because of the particular placement and mastery tests taken. At each grade level, there was a random selection of 100 tests for each test length, but the sizes of the item pools differed. The sizes of the item pools follow:

Grade 3:	Multiplication	22 items
Grade 5:	Multiplication	26 items
Grade 5:	Division	17 items
Grades 6-8:	Multiplication	26 items
Grades 6-8:	Division	17 items

For each test length of heterogeneous items, all three reliability estimates were computed. Coefficient kappa was computed using data from heterogeneous pre-tests and post-tests while the estimate of kappa ($\hat{k}_H$) and the coefficient of agreement ($p_{c(s)}$) were computed using the pre-test data only. In cases where there was more than one test of a particular length, mean estimates and standard errors were computed.

Each reliability estimate was computed for each heterogeneous test constructed. Each estimate was examined to see how the violation of item homogeneity in tests of constant length will affect the estimates. The estimates for the heterogeneous tests were compared to the estimates for the homogeneous tests of the same length. Secondly, the estimate of k was compared to the standard, kappa, to see how they behaved relative to kappa for tests which are not homogeneous.

F. Validation

A different yet related aspect of this research concerns the question of validity of the mastery decisions. A random sample of students in Grades 5 and 7 was selected for a validation study of the mastery decisions. Students in Grades 5 and 7 in the MCPS take the Iowa Test of Basic Skills (ITBS) as part of the regular instructional program. Therefore, the scores in the Math Computation section of the ITBS of the students selected were compared to the results of the mastery decisions. If, for example, a group of students has been consistently classified as masters and these classifications have been determined to be reliable, do the students' scores on the ITBS support the classifications? The ITBS was used as an outside criterion of proficiency to try to examine the issue of how a reliable mastery classification relates to a measure of proficiency which exists outside the system. In classical test theory, validity is bounded by the square root of reliability, but it is unclear whether any such relationship exists for criterion-referenced tests. In order to examine the relationship between

reliability and validity for the ISM criterion-referenced tests, the students' total scores on the ITBS mathematics section were correlated with their scores on the different ISM mastery tests. These correlations were then compared to the reliability estimates for the tests of different lengths to try to determine if a relationship exists for these tests, and consequently whether a relationship could be specified between reliability and validity for criterion-referenced tests.

The validation analysis consisted of computing Pearson Product Moment coefficients between the scores of students in Grades 5 and 7 on the Total Math section of the ITBS and the mastery tests, as given by MCPS, which these students took. For each grade level there were two mastery tests taken. Thus, there was a total of four validity coefficients computed.

G. Relationship Between the KR21 and the Criterion-Referenced Reliability Coefficients

One final aspect of this research involved the comparison of the criterion-referenced reliability estimates to norm-referenced or classical reliability coefficients computed from the same ISM test data. The first analysis consisted of computing KR21 coefficients for the different length tests and then computing a PPM correlation between the KR21 coefficients and each of the three criterion-referenced coefficients computed. This analysis addressed the issue of the relationship between internal consistency and the criterion-referenced indices.

Kuder-Richardson (internal consistency) coefficients were computed for all five objectives. They were computed for total test length for each objective as well as for each test length which was analyzed for the test length analysis. This means, for example, that for Grade 6, DI 10-N, a KR coefficient was computed for the total test length of 9 items, for a random sample of 7 items and for a random sample of 5 items. These coefficients for the different test lengths were then compared to the kappa coefficients for the same 5 and 7 items, respectively. The Kuder-Richardson coefficients for the different test lengths were also compared to the Huynh and Subkoviak coefficients for the same test items.

Comparison Between Test-Retest Coefficients and the Criterion-Referenced Reliability Coefficients

A second comparison between norm-referenced and criterion-referenced coefficients was undertaken by computing test-retest coefficients between the two test administrations. These test-retest coefficients were computed for all five objectives and were computed by grade level for the objectives for Grades 6-8. These test-retest coefficients were then compared to the relevant criterion-referenced coefficients by means of a Pearson Product Moment correlation. It was possible to do this analysis because, as was explained at the beginning of the Methodology section, there were two test administrations, separated by a two-week period. It is especially interesting to note the relationship between test-retest coefficients and the two reliability estimates of Huynh (1976) and Subkoviak (1976) because there are actual data from a

second test administration and it will be interesting to see what the two simulations produce for second forms.

The initial computation of the test-retest coefficients consisted of computing the coefficients on the mastery tests, as they are given by MCPS. Subsequently, test-retest coefficients were computed on a random sample of 7 items and on the total items for each objective. These test-retest coefficients were then compared to the kappa, Huynh, and Subkoviak coefficients for the same items. The test-retest coefficients for 3 items and 7 items were computed on the same items as the Kuder-Richardson internal consistency coefficients to allow for a comparison of these coefficients.

Chapter 5

RESULTS

In this research there were eight analyses performed. Six analyses examined the effect of different variables on the three reliability coefficients k, $\hat{k}_H$, and $p_{c(s)}$ and the relationship of these coefficients to each other. These six analyses were: (1) the effect of test length, (2) the combined effect of cut-off score and test length, (3) the combined effect of ability grouping and cut-off score, (4) the effect of sample size, (5) the effect of heterogeneity of test content, (6) the validity of the criterion-referenced tests. The last two analyses examined: (1) the relationship between two norm-referenced coefficients (KR21 and test-retest) and the criterion-referenced coefficients, and (2) the relationship between the empirical standard, k, and the single administration criterion-referenced coefficient intended to estimate kappa, $\hat{k}_H$. Each analysis is discussed in a separate section.

In each table the mean values for a particular test length and the standard errors are both reported. The mean values of the coefficients were determined by: (1) constructing all possible tests of a given length from the item pool (see Equation 17, p. 60); (2) computing all three reliability coefficients for each combination; (3) finding the mean of each coefficient over all the test combinations. When the value of the coefficient is reported for the maximum number of items for an objective, the coefficient is simply a single value and thus there is no standard error.

It should be noted that in each table in the Results there are two sets of results reported for $P_{c(s)}$. One set of results is based on the assumption of a simple binomial model and one set of results is based on the assumption of a compound binomial model (see Chapter 4, pp. 57-58).

I. Test Length

The first analysis conducted examined the effect of test length on the three reliability coefficients. The test items measuring one objective were reorganized to construct tests of different lengths using the method described in the preceding paragraph. Mastery level was held constant at = 80 percent for all test lengths and all grade levels. In Table 2, the mean values for k, $\hat{k}_H$ and $P_{c(s)}$ are presented for test lengths of n = 5, 7, 8, 9, 10, 13. The averaging is over grade level and over all the different test combinations. Results by grade level can be found in the Appendix, Tables 25-29. In addition to the average coefficients, average standard errors are presented. Within each grade level a standard error is generated by considering the reliability values for all possible tests of a given length. The number of such tests can be computed by the combination formula (see Equation 17, p. 60). In Table 2, the standard errors are averaged over grade.

The predicted results of the test length analysis were only partially attained.

TABLE 2

The effect of test length (N) on the expected value and s.e. of the three reliability coefficients

Test Length	E(k)	s.e. k	$E(\hat{k}_H)$	s.e. $\hat{k}_H$	$E(p_{c(s)})$ (simple)	s.e. $p_{c(s)}$ (simple)	$E(p_{c(s)})$ (compound)	s.e. $p_{c(s)}$ (compound)
5 items	.375	.091	.513	.057	.709	.039	.888	.042
7 items	.388	.078	.505	.032	.699	.022	.877	.024
8 items	.511	-	.559	-	.754	-	.890	-
9 items	.335	-	.429	-	.724	-	.956	-
10 items	.450	-	.725	-	.826	-	.958	-
13 items*	.354	-	.796	-	.802	-	.954	-
Average	.402	.085	.588	.045	.752	.031	.921	.033

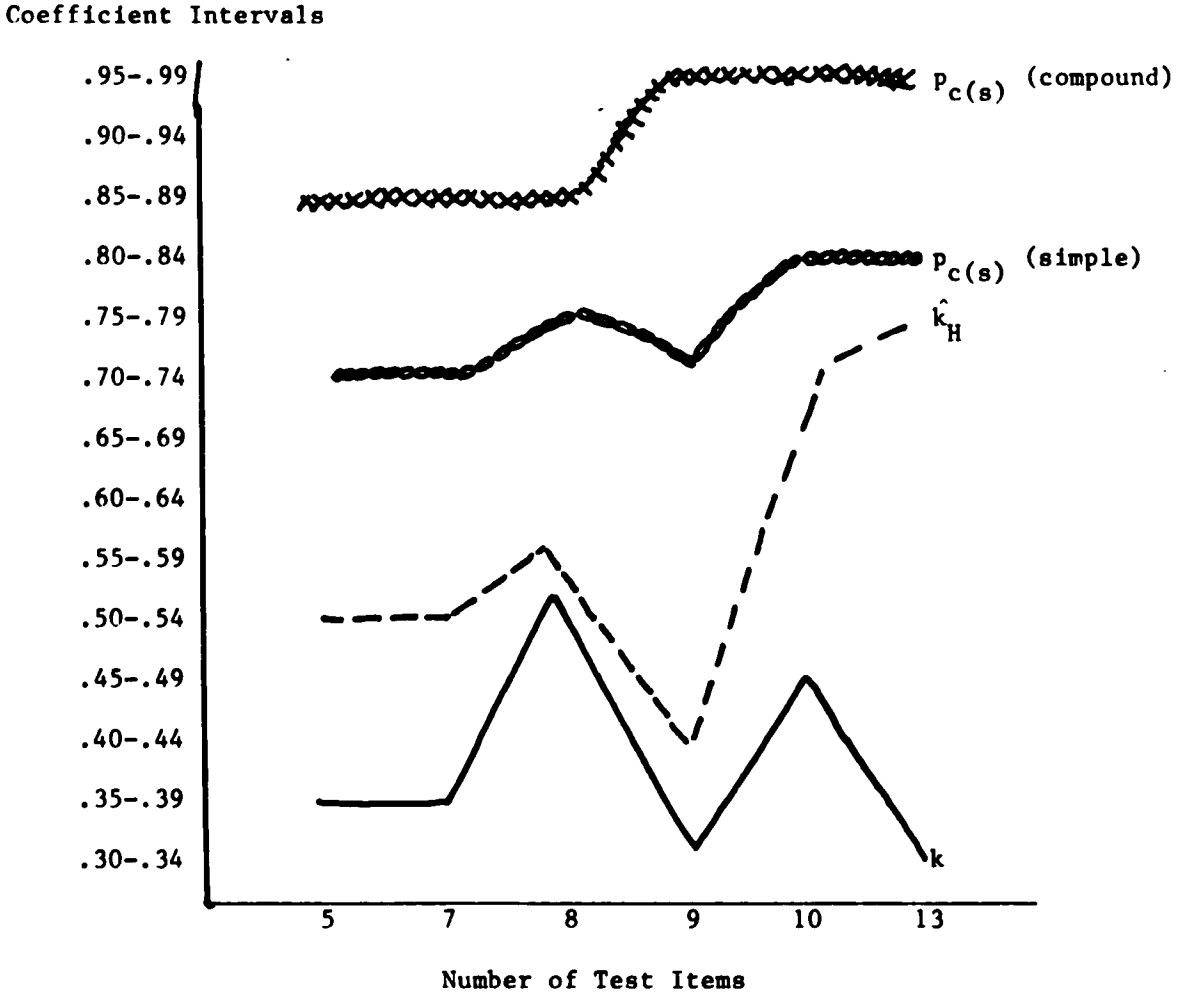
* - Not mean values because only one group took this test length.

As can be seen in Table 2, the general increase in the size of the coefficients by test length is true only for the increase from 5 to 8 items. All the coefficients except $p_{c(s)}$ (compound) are particularly low for the 9-item tests. When the tests are lengthened beyond 9 items there is no clear pattern of general increase for any of the coefficients. Coefficient k in fact decreases markedly for the 13-item test. However, for all the coefficients except k an increasing pattern for the reliability coefficients as the test is lengthened may not be detectable due to the size of the standard errors. These patterns may be seen in Figure 1.

With regard to the 9-item tests, the same sample of students took these items measuring multiplication (MU08-L: multiplication of 3- and 4-digit numbers by 2-digit numbers) as took the division items (DI10-N: division of 3- and 4-digit numbers by 2-digit numbers). The distribution of the coefficients for the division items is quite different (see Appendix, Tables 26 and 27). For the division objective, the majority of coefficients across the test lengths and across the grade levels is .40 or above while for the multiplication objective all but one of the ks are below .40, and all but one of the $\hat{k}_H$s are .40 and below. The values of $p_{c(s)}$ (simple binomial) also are at their lowest - .60 to .70 - for this multiplication objective in the test length analysis. This distribution would suggest that the multiplication items are not good measures or that the sample of students is not adequately competent in multiplication to perform consistently on a repeated measure of performance in multiplication.

FIGURE 1

Distributions of mean coefficient values of k, $\hat{k}_H$, $p_{c(s)}$ (simple) and $p_{c(s)}$ (compound) for tests of different lengths



Further, in Table 2, we note the great difference among the standard errors for the coefficients. The standard errors for k were comparatively large (over .06) while the standard errors for $\hat{k}_H$ were moderate to low and those for $p_{c(s)}$ were extremely low.

With respect to size of the reliability coefficients, in almost all instances, the size of the $\hat{k}_H$ coefficients was greater than the size of the k coefficients. The coefficient of agreement from both the simple and compound binomial models was extremely high, .60 to .79 for the simple model and over .80 for the compound model, in contrast to the .20-.59 range for most of the k and $\hat{k}_H$ coefficients. The size of the coefficients of agreement would be expected to be greater than the size of k or $\hat{k}_H$ because of there being no correction for chance agreement for $p_{c(s)}$.

An examination of rank order correlations between k, the standard, and each of the three other coefficients provides an interesting comparison among the coefficients. The rank order correlation between k and $\hat{k}_H$ is .315; the rank order correlation between k and $p_{c(s)}$ (simple) is .20 and the correlation between k and $p_{c(s)}$ (compound) is -.143. Thus, the agreement is poor.

The relationship between k and $\hat{k}_H$ - i.e., whether $\hat{k}_H$ is a good estimate of k - can be examined by determining the difference scores between the two coefficients for the different test lengths. In Table 3 are presented the mean difference scores and ranges of the two coefficients by test length.

TABLE 3

Ranges of the k and $\hat{k}_H$ coefficients and mean difference scores between the coefficients for different test lengths

Test Length	Range of k	Range of $\hat{k}_H$	$\bar{d}$
5 items	.211-.522	.205-.706	-.082
7 items	.267-.492	.269-.735	-.122
8 items	.338-.629	.397-.644	-.004
9 items	.225-.474	.306-.529	-.094
10 items	.468-.525	.724-.778	-.255
13 items*	E(k)=.354	$E(\hat{k}_H)$=.796	-.426

* - Not a mean difference because only one group took this test length.

It is evident from examining Table 3 that $\hat{k}_H$ seems to be a fairly good estimate of k for the shorter test lengths, but for the test lengths of 10- and 13-item tests $\hat{k}_H$ is not a good estimate of k. Additionally the fact that the mean differences between k and k_H are all negative, may indicate that $\hat{k}_H$ is a biased estimated of k.

II. The Combined Effect of Cut-off Score and Test Length

The second analysis conducted examined the combined effect of cut-off score and test length. Each analysis was conducted by grade level since each grade took different objectives. Results of the grade level analysis can be seen in the Appendix, Tables 30-40. Since the results by grade level did not show any discernible pattern the results were collapsed across grade level to look at the combined effect of cut-off score and test length.

Since in many cases it was not possible to figure exact percentages for cut-off scores of 60 and 80 percent, items were selected to come as close as possible to the specified cut-off scores. The following table demonstrates the exact percentages of items selected for particular test lengths:

TABLE 4

Number of items and percentage mastery for different test lengths used in the analysis of the combined effect of cut-off score and test length

Number of Items in Test	Prespecified Mastery Level	Number of Items and Percentage Mastery
5	60%	3(60%)
	80%	4(80%)
	100%	5(100%)
7	60%	4(57%)
	80%	6(85%)
	100%	7(100%)
8	60%	5(67%)
	80%	6(75%)
	100%	8(100%)
9	60%	6(67%)
	80%	7(78%)
	100%	9(100%)
10	60%	6(60%)
	80%	8(80%)
	100%	10(100%)
13	60%	8(61%)
	80%	11(84%)
	100%	13(100%)

Table 5 presents mean values and standard errors for all the coefficients by test length and cut-off score. As with the data in the test length analysis, the averaging is over grade levels and over all the different test combinations. In addition, average standard errors

TABLE 5

The effect of cut-off score and test length on the expected value and s.e. of the three reliability coefficients

Cut-off Score	Test Length	E(k)	s.e. k	$E(\hat{k}_H)$	s.e. $\hat{k}_H$	$E(p_{c(s)})$ (simple)	s.e. $p_{c(s)}$ (simple)	$E(p_{c(s)})$ (compound)	s.e. $p_{c(s)}$ (compound)
60%	5 items	.397	.182	.433	.176	.830	.051	.945	.053
	7 items	.411	.195	.482	.178	.863	.046	.964	.038
	8 items	.487	.085	.552	.117	.817	.023	.938	.046
	9 items	.292	.118	.405	.106	.832	.044	.982	.012
	10 items	.601		.756		.903		.971	
	13 items*	.703		.809		.927		.995	
	Average	.481	.145	.572	.144	.862	.041	.968	.037
80%	5 items	.375	.097	.456	.153	.710	.053	.888	.064
	7 items	.388	.068	.505	.140	.699	.068	.877	.058
	8 items	.511	.107	.559	.098	.754	.023	.890	.046
	9 items	.336	.108	.430	.092	.724	.035	.956	.028
	10 items	.497		.760		.826		.958	
	13 items*	.354		.796		.802		.954	
	Average	.410	.095	.584	.121	.753	.045	.921	.049
100%	5 items	.269	.049	.416	.137	.780	.081	.743	.073
	7 items	.259	.086	.443	.132	.733	.073	.763	.070
	8 items	.193	.181	.458	.068	.795	.071	.778	.077
	9 items	.333	.120	.354	.078	.730	.049	.750	.027
	10 items	.196		.670		.788		.890	
	13 items*	.264		.716		.806		.892	
	Average	.252	.109	.509	.104	.772	.069	.802	.062

* - Not mean values because only one group took this test length.

are presented. With in each grade level a standard error is generated by considering the reliability values for all possible tests of a given length.

In general, the mean values of k and $\hat{k}_H$ are highest at the 60 and 80 percent cut-off scores with a decline, rather precipitous for k, at the 100 percent cut-off score. The standard errors for the k and $\hat{k}_H$ coefficients are all above .06.

The mean values of the two $p_{c(s)}$ coefficients are much higher than the means of the k or $\hat{k}_H$ coefficients, at all cut-off scores. For both $p_{c(s)}$ coefficients the maximum mean value occurs at the 60 percent cut-off score while the lowest mean value for $p_{c(s)}$ (simple) is at the 80 percent cut-off and that for $p_{c(s)}$ (compound) is at the 100 percent cut-off. The standard errors for both $p_{c(s)}$ coefficients are considerably lower than those for k and $\hat{k}_H$.

In examining the effect of test length within each cut-off score, there appears to be a general increase in the size of the coefficients as the test is lengthened. However, there is no consistent pattern. As was the case with the test length analysis a consistent pattern of increase may be being obscured by the size of the standard errors. Also consistent with the test length analysis is the smaller values of the coefficients for the 9-item tests at each cut-off score.

In order to examine the relationship between k and the two single administration coefficients rank order correlations were computed between k and $\hat{k}_H$ and between k and $p_{c(s)}$ (simple and compound) within each cut-off score. Thus the coefficients for the different test lengths were rank ordered within each cut-off score and rank order correlation coefficients were computed.

The greatest positive relationship between k and the other coefficients occurs at the 60 percent cut-off score. There is a perfect 1.0 correlation between k and $\hat{k}_H$ at this cut-off score. The correlation between k and $p_{c(s)}$ (simple) is .60 while that between k and $p_{c(s)}$ (compound) is .26.

There is virtually no agreement between k and the other coefficients at either the 80 percent or the 100 percent cut-off score. At the 80 percent cut-off score the correlation between k and $\hat{k}_H$ is .315 and those between k and $p_{c(s)}$ (simple) and k and $p_{c(s)}$ (compound) are .20 and -.143, respectively. At the 100 percent cut-off score, the rank order correlations are: -.02 between k and $\hat{k}_H$ and -.54 between k and both $p_{c(s)}$ (simple and $p_{c(s)}$ (compound).

If the coefficients are averaged over test length by cut-off score, k and $p_{c(s)}$ (compound) both are highest at the 60 percent cut-off and decline steadily to the 100 percent cut-off. Neither $\hat{k}_H$ nor $p_{c(s)}$ (simple) follows the same pattern. These results which confirm the lack of consistent patterns in the correlational analysis are presented below in Table 6.

In an attempt to address the issue of whether $\hat{k}_H$ is a good estimate of k, average difference scores were computed between k and $\hat{k}_H$ for each test length and for each cut-off score. The results of these computations may be seen in Table 7.

At the 60 percent and at the 80 percent cut-off scores, $\hat{k}_H$ appears to be a reasonable estimate of k for the shorter test lengths. At the 100 percent cut-off score, however, $\hat{k}_H$ seems to be a reasonable estimate of k only for the test lengths of 8 items and 9 items. Thus there is no consistent pattern across all cut-off scores for how well $\hat{k}_H$ estimates k. Additionally in this analysis as was the case for the test length analysis, the average difference scores are all negative, indicating that $\hat{k}_H$ is consistently overestimating k.

TABLE 6

The effect of cut-off score, averaged over test length on the expected value and s.e. of the three reliability coefficients

Cut-off Score	E(k)	s.e.(k)	$E(\hat{k}_H)$	$s.e.(\hat{k}_H)$	$E(p_{cs}s)$	$s.e.(p_{cs}s)$	$E(p_{cs}c)$	$s.e.(p_{cs}c)$
60%	.481	.145	.572	.144	.862	.041	.968	.037
80%	.410	.095	.584	.121	.753	.045	.921	.049
100%	.252	.109	.509	.104	.772	.069	.802	.062

TABLE 7

Ranges of the k and $\hat{k}_H$ coefficients and mean difference scores between the coefficients for different test lengths within three cut-off scores

Cut-off Score	Test Length	Range of k	Range of $\hat{k}_H$	$\bar{d}$
60%	5 items	.112-.723	.149-.703	-.036
	7 items	.109-.754	.174-.743	-.071
	8 items	.393-.593	.359-.657	-.064
	9 items	.119-.381	.260-.514	-.113
	10 items	.459-.742	.728-.783	-.155
	13 items*			d*=-.106
80%	5 items	.211-.522	.205-.706	-.082
	7 items	.266-.492	.269-.735	-.118
	8 items	.338-.629	.397-.644	-.047
	9 items	.225-.474	.306-.529	-.099
	10 items	.468-.525	.724-.778	-.255
	13 items*			d*=-.442
100%	5 items	.153-.332	.205-.664	-.146
	7 items	.096-.403	.239-.685	-.184
	8 items	-.089-.384	.342-.511	-.121
	9 items	.175-.360	.266-.453	-.02
	10 items	.132-.259	.635-.704	-.474
	13 items*			d*=-.452

* - Indicates only one case of the test length and thus d is actual difference between k and $\hat{k}$.

III. The Combined Effect of Ability Level and Cut-off Score

The analysis of the effect of ability level on the three coefficients consisted of conducting the analyses of test length and cut-off score within two ability groups, high and low, for each objective. Scores from the Iowa Test of Basic Skills (ITBS) were used to identify students of high ability and low ability in mathematics. The criterion level of the 50th percentile based on the norming sample was chosen for the identification process. Those students whose total Math Score on the ITBS was above the 50th percentile were considered high ability while those students whose Total Math Score was in the 50th percentile or below were considered low ability.

It would have been more desirable to have been able to make finer distinctions in the ability groupings such as students with scores in stainines 1, 2, 3 for low ability and students with scores in stanines 7, 8, 9 for high ability. However, the particular sample of students is one of relatively high achievement and there were not enough students with math achievement scores in stanines 1, 2, 3 to make up a low ability group. Even using the criterion of the 50th percentile to distinguish between high and low ability, the analysis could not be done for the third grade sample because there were only five students who met the criterion for low ability. In the fifth grade there were 26 students in the low ability group and 28 in the high ability group. For grades 6-8, the low ability group consisted of 43 students while the high ability group was three times as large, N = 129. Thus the total for the low ability group was 69 while the total for the high ability group was 157.

Table 8 presents mean values and standard errors for al the coefficients by cut-off score averaged over test length, for both the low ability and the high ability groups. The averaging is first over grade levels and over all the different test combinations (see Equation 17, p. 60). Then the test length results were collapsed by cut-off score. The results were collapsed over grade level because there was no consistent pattern by grade level for either ability group. Results of the grade level analysis can be found in the Appendix, Tables 41-48. In addition, average standard errors are presented. Within each grade level a standard error is generated by considering the reliability values for all possible tests of a given length.

TABLE 8

The combined effect of ability level and cut-off score on the expected value and s.e. of the three reliability coefficients for low ability and high ability students

Cut-off Score	Ability Level	E(k)	s.e.(k)	$E(\hat{k})$	s.e.$(\hat{k})$	$E(p_{c(s)})$ (simple)	s.e.$(p_{c(s)})$ (simple)	$E(p_{c(s)})$ (compound)	s.e.$(p_{c(s)})$ (compound)
60%	Low	.473	.121	.617	.041	.815	.026	.936	.028
	High	.346	.096	.471	.046	.845	.023	.967	.017
80%	Low	.487	.105	.608	.042	.779	.028	.885	.029
	High	.381	.093	.494	.043	.726	.027	.905	.032
100%	Low	.134	.199	.509	.047	.822	.020	.850	.031
	High	.163	.109	.424	.046	.718	.022	.757	.033

An examination of Table 8 reveals that only $\hat{k}_H$ is consistently somewhat higher for the low ability group. The size of the standard errors for k are such that there is not much difference between the values of k for the two ability groups. Neither $p_{c(s)}$ (simple nor $p_{c(s)}$ (compound) follows a consistent pattern of being higher or lower for one ability group.

Figure 2 presents the distributions of the numbers of each reliability coefficient in coefficient intervals for both the low and the high ability groups. There was a total of 12 of each coefficient computed for each ability group and cut-off score.

It is clear from the figures that there is very little overlap between the coefficients of agreement and k and $\hat{k}_H$ for either ability group.

In order to investigate whether $\hat{k}_H$ is a good estimate of k for the two ability groups mean difference scores were computed between k and $\hat{k}_H$ by test length within each cut-off score for each ability groups. This analysis could only be done for the test lengths of 5-items, 7-items and 8-items because there was only one instance of the test lengths of 9-items, 10-items and 13-items. The data from these computations are presented in Table 9.

FIGURE 2

Distributions of the numbers of the three reliability coefficients in coefficient intervals for low ability and high ability students

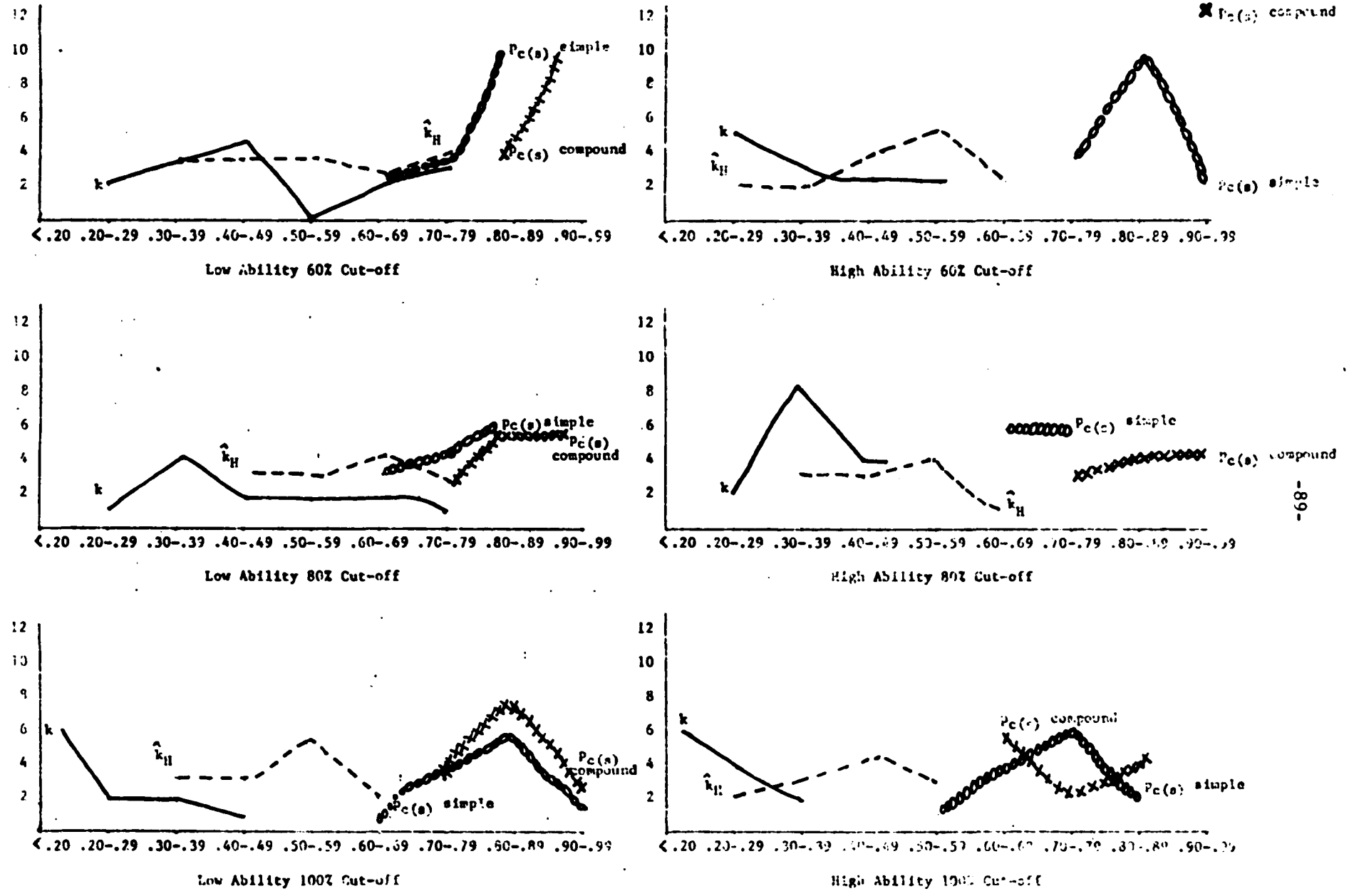


TABLE 9

Ranges of the k and $\hat{k}_H$ coefficients and mean difference scores between the coefficients by test length and cut-off score for the low ability and high ability groups

Cut-off Score	Test Length	Low Ability			High Ability		
		Range of k	Range of $\hat{k}_H$	$\bar{d}$	Range of k	Range of $\hat{k}_H$	$\bar{d}$
60%	5 items	.249-.668	.428-.671	-.116	.245-.526	.224-.516	-.04
	7 items	.267-.719	.488-.715	-.147	.257-.525	.266-.577	-.09
	8 items	.394-.752	.529-.728	-.056	.304-.401	.453-.567	-.158
80%	5 items	.314-.697	.435-.654	-.110	.289-.399	.279-.542	-.092
	7 items	.273-.688	.468-.682	-.143	.324-.371	.345-.598	-.132
	8 items	.518-.738	.520-.709	.014	.413-.419	.478-.563	-.105
100%	5 items	.118-.436	.352-.593	-.196	.147-.276	.206-.509	-.136
	7 items	-.012-.336	.365-.611	-.324	.004-.318	.298-.547	-.259
	8 items	-.040-(-.086)	.369-.576	-.535	-.139-.275	.404-.433	-.351

These data would indicate that $\hat{k}_H$ consistently somewhat overestimates k at the 60 and 80 percent cut-off scores and there is little difference in the amount of overestimation between the two ability groups. At the 100 percent cut-off score, $\hat{k}_H$ greatly overestimates k for both ability groups.

IV. Sample Size

The fourth analysis examined the effect of sample size on the reliability estimates. The analysis was conducted by objective taken which also means by grade level. In the case of the two objectives taken by Grades 6-8, the students in Grade 7 were selected for this analysis. The reason for this is that there were approximately equivalent numbers of students in Grade 7 as in Grades 3 and 5. The analyses for effect of sample size entailed examining the different test lengths and cut-off scores for a sample of one classroom ($N \tilde{=} 25$) for each objective and then for two classrooms ($N \tilde{=} 50$). The data from the analysis of single classrooms and two classrooms are presented in Table 10. In this table the mean values are presented for the three cut-off scores of 60 percent, 80 percent and 100 percent averaged first over grade levels and over all the different test combinations. These results are then collapsed over test length. In addition average standard errors are presented. Within each grade level a standard error is generated by considering the reliability values for all possible tests of a given length (see Equation 17, p. 60) and the standard errors are averaged over grade. They are further averaged over test length. Grade level results are available in the Appendix, Tables 49-53 (one classroom samples and Tables 30, 31, 32, ,34, 38 (two classroom samples).

TABLE 10

The effect of sample size on the expected value and s.e. of the three reliability coefficients by cut-off score for one classroom (N = 25) and two classroom (N = 50) samples

Cut-off Score	Sample Size	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_{c(s)})$ (simple)	$s.e.(p_{c(s)})$ (simple)	$E(p_{c(s)})$ (compound)	$s.e.(p_{c(s)})$ (compound)
60%	One class	.425	.140	.533	.050	.894	.019	.984	.010
	Two classes	.511	.081	.551	.066	.879	.019	.975	.015
80%	One class	.378	.105	.548	.052	.794	.029	.954	.004
	Two classes	.385	.072	.565	.038	.768	.022	.940	.026
100%	One class	.064	.117	.495	.056	.734	.020	.803	.038
	Two classes	.242	.089	.501	.040	.739	.015	.803	.036

An examination of Table 10 shows that the mean values of the three coefficients seem to change very little from the one classroom ($N \doteq 25$) to the two classroom ($N \doteq 50$) sample. The sizes of the k and $\hat{k}_H$ coefficients appear slightly larger in the two classroom sample while the coefficients of agreement appear to be slightly larger for the one classroom sample.

With respect to the size of the standard errors of the coefficients, as predicted, the size of the mean values of the standard errors for k is somewhat smaller in the two classroom sample, at all cut-off scores. For $\hat{k}_H$ at the 60 percent cut-off score, the size of the mean standard error is smaller for the one classroom samples, while at the 80 and 100 percent cut-off scores the mean standard errors for the two classroom sample are smaller. For the coefficients of agreement the size of the standard errors is virtually the same for the two samples sizes except for that for $p_{c(s)}$ (compound) at the 80 percent cut-off which is much smaller for the one classroom sample. However, the standard errors for $p_{c(s)}$ (compound) are higher than those for $p_{c(s)}$ (simple) for both sample sizes and the range of the standard errors is greater in the two classroom sample.

In order to investigate $\hat{k}_H$'s estimation of k, mean difference scores between k and $\hat{k}_H$ were computed for different test lengths by cut-off score for the one classroom and two classroom samples. The difference scores for the 9- and 13-item tests are absolute differences because there was only one of each test for the samples. The results of these computations can be seen in Table 11.

TABLE 11

Ranges of the k and k_H coefficients and mean difference scores by cut-off score and test length for one classroom (N = 25) and two classroom (N = 50) samples

Cut-off Score	Test Length	One Classroom			Two Classrooms		
		Range of k	Range of $\hat{k}_H$	$\bar{d}$	Range of k	Range of $\hat{k}_H$	$\bar{d}$
60%	5 items	.105-.659	.135-.736	-.072	.184-.723	.149-.703	-.015
	7 items	.000-.681	.143-.772	-.127	.125-.754	.174-.743	-.037
	8 items	.000-.559	.344-.605	-.195	.393-.593	.359-.657	-.015
	9 items	.353	.218	d =.135	.119	.514	d*=-.395
	10 items	.500-.657	.696-.828	-.184	.459-.742	.728-.783	-.155
	13 items*	.606	.828	d =-.222	.703	.809	d*=-.106
80%	5 items	.174-.498	.192-.722	-.09	.272-.522	.205-.706	-.089
	7 items	.206-.517	.244-.743	-.162	.266-.492	.269-.735	-.141
	8 items	.338-.600	.392-.609	-.029	.338-.554	.397-.644	-.075
	9 items	.358	.271	d = .087	.258	.529	d*=-.271
	10 items	.395-.404	.709-.788	-.349	.468-.525	.724-.778	-.255
	13 items*	.324	.799	d =-.475	.354	.796	d*=-.442
100%	5 items	.089-.389	.204-.659	-.214	.153-.332	.205-.664	.179
	7 items	.024-.387	.229-.674	-.295	.096-.403	.239-.685	-.245
	8 items	-.185-.087	.371-.512	-.491	-.089-.134	.342-.511	-.404
	9 items	.355	.259	d =.096	.175	.453	d*=-.278
	10 items	-.008-.333	.647-.688	-.505	.132-.259	.635-.704	-.474
	13 items*	.114	.696	d =-.582	.264	.716	d*=-.452

*Indicates only one case of test length and so d is actual difference between k and $\hat{k}_H$.

At the 60 percent cut-off score $\hat{k}_H$ is a better estimate of k for the larger (e.g., two classroom) sample for the shorter test lengths. At the 80 percent cut-off, the mean difference scores between the coefficients are about the same for the two sample sizes and the difference scores are higher than those for the 60 percent cut-off score. At the 100 percent cut-off score the mean differences between the coefficients are very large for both sample sizes. Thus only for the shorter test lengths at the 60 percent cut-off score is $\hat{k}_H$ a fairly good estimate of k. It should be noted, however, that consistent with the previous analyses $\hat{k}_H$ is an overestimate at almost all test lengths and cut-off scores.

V. Heterogeneity of Test Content

The analysis of item heterogeneity consisted of constructing tests of different lengths using items from several different objective levels. The test lengths examined were the same as those examined for the test length analysis (e.g., 5, 7, 8, 9, 10 and 13 items) and the cut-off score was held constant at 80 percent. The content of the test was the only thing which varied.

It was possible to vary the content of the tests because the placement tests which the students took consisted of items measuring several different objectives (or skills) in multiplication and division each objective being measured by a few items. Therefore these items were organized into 2-item pools, multiplication and division. The combination program was run (see Equation 17, p. 60) for each pool of items and for each test length of each objective 100 combinations of items were selected at random for examination. Grade level and skill

area multiplication and division results can be seen in the Appendix, Tables 54-62.

In Table 12, the mean values for k, $\hat{k}_H$ and $p_{c(s)}$ are presented for test lengths of N = 5, 7, 8, 9, 10, 13 for homogeneous and heterogeneous tests. The averaging is over grade levels and over the 100 different test combinations. In addition, average standard errors are presented. Within each grade level a standard error is generated by considering the reliability values for all possible tests of a given length (see Equation 17, p. 60). In Table 12, the standard errors are averaged over grade.

The results for the heterogeneous tests are different from the results of this analysis with homogeneous test content. When the test content was homogeneous, there was a general increase in the size of the mean values of the coefficients up to a test length of 8 items, beyond which the pattern of a general increase was not discernible.

As can be seen in Table 12, the mean values of k and $\hat{k}_H$ for the heterogeneous tests do not show a pattern of increase as the test is lengthened, while the maximum mean value of k occurs for a test length of 10 items and the maximum mean value of $\hat{k}_H$ occurs for a test length of 13 items. Additionally, the size of the standard errors for k and $\hat{k}_H$ are higher for the heterogeneous tests. The mean values of the $p_{c(s)}$ coefficients show much less variability in behavior between the two types of tests.

Figure 3 presents the distributions of the coefficients for the homogeneous and the heterogeneous tests.

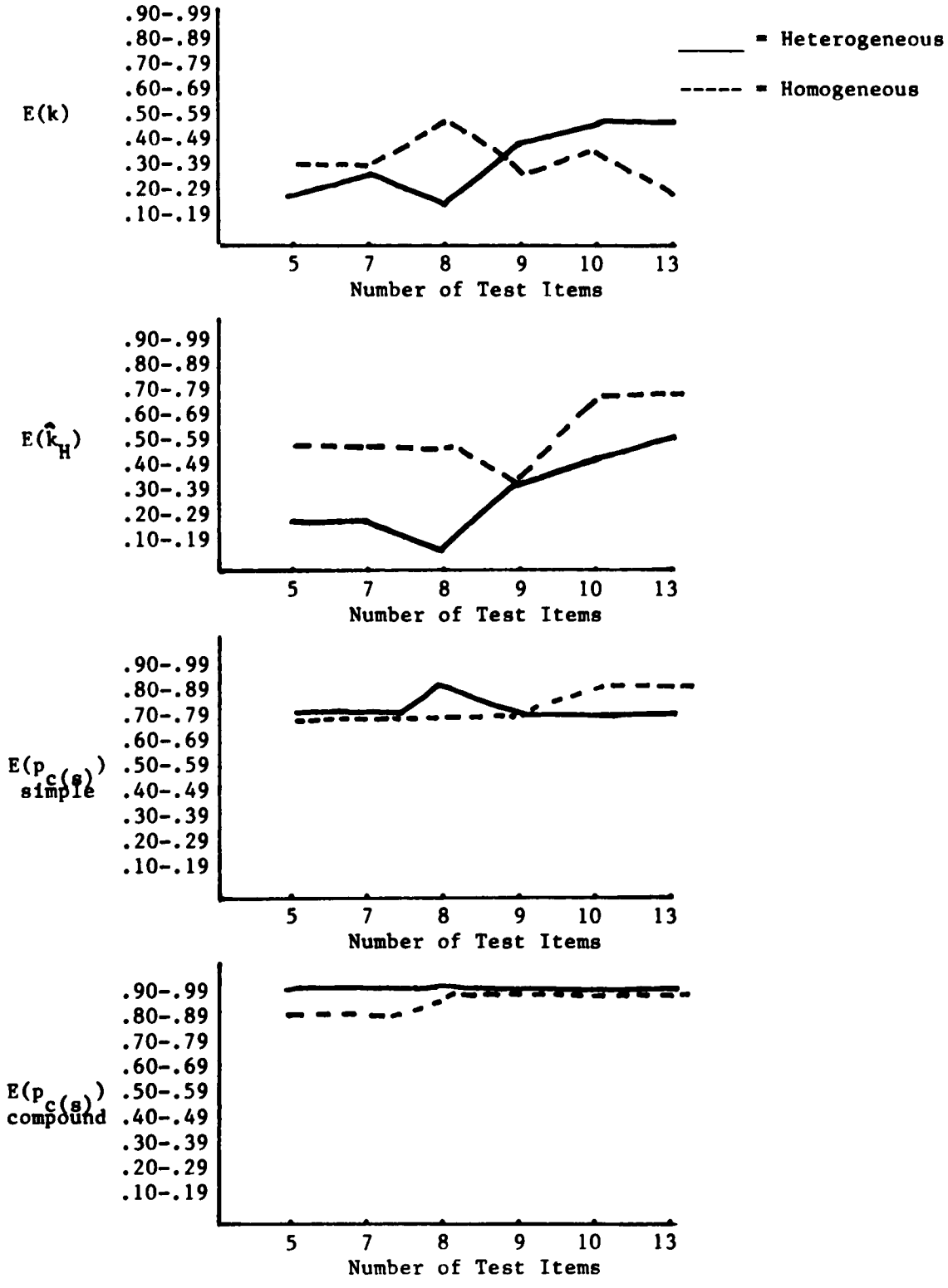
TABLE 12

The effect of test length on the expected value and s.e. of the reliability coefficients for tests of homogeneous and heterogeneous test content

Test Length	Test Content	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_{c(s)})$ (simple)	$s.e.(p_{c(s)})$ (simple)	$E(p_{c(s)})$ (compound)	$s.e.(p_{c(s)})$ (compound)
5 items	Homogeneous	.375	.091	.513	.057	.709	.039	.888	.042
	Heterogeneous	.295	.154	.260	.119	.789	.092	.976	.028
7 items	Homogeneous	.388	.078	.505	.032	.699	.022	.877	.024
	Heterogeneous	.323	.136	.296	.102	.744	.087	.969	.022
8 items	Homogeneous	.511	-	.559	-	.754	-	.890	-
	Heterogeneous	.230	.194	.157	.104	.854	.132	.997	.006
9 items	Homogeneous	.335	-	.429	-	.724	-	.956	-
	Heterogeneous	.423	.115	.415	.062	.795	.044	.975	.014
10 items	Homogeneous	.450	-	.725	-	.826	-	.958	-
	Heterogeneous	.570	.102	.566	.082	.761	.037	.937	.019
13 items	Homogeneous	.354	-	.796	-	.802	-	.954	-
	Heterogeneous	.520	.081	.622	.052	.762	.024	.967	.014

FIGURE 3

Distributions of k, $\hat{k}_H$ and $p_{c(s)}$ for tests composed of heterogeneous and tests of homogeneous test content



It can be seen from Figure 3 that the distributions of mean values of the coefficients are almost the same for the two test types for the $p_{c(s)}$ coefficients while they are different for k and $\hat{k}_H$. For k not only is the pattern different, but the coefficients are higher for the heterogeneous test content. For $\hat{k}_H$ the coefficient behaves similarly over the two types of tests, but the mean values are quite a bit higher for the homogeneous test content.

The rank order correlation coefficients computed between k and each of the other reliability coefficients show a strong relationship between k and $\hat{k}_H$ and no agreement between k and $p_{c(s)}$ (simple) or between k and $p_{c(s)}$ (compound). The rank order correlation between k and $\hat{k}_H$ is .944. The correlation between k and $p_{c(s)}$ (simple) is -.58 while that between k and $p_{c(s)}$ (compound) is -.89. For the homogeneous test content there was no agreement between k and the other coefficients as determined by the rank order correlation coefficients. Those coefficients were: .315 (k and $\hat{k}_H$); .20 (k and $p_{c(s)}$ simple); and -.143 (k and $p_{c(s)}$ compound).

Mean difference scores were computed between k and $\hat{k}_H$ for the different test lengths for the heterogeneous tests. These difference scores and ranges of the two coefficients by test length for both types of tests, homogeneous and heterogeneous, are presented in Table 13.

TABLE 13

Ranges of the k and $\hat{k}_H$ coefficients and mean difference scores between the coefficients by test length and cut-off score for the low ability and high ability groups

	Test Length	Homogeneous			Heterogeneous		
		Range of k	Range of $\hat{k}_H$	$\bar{d}$	Range of k	Range of $\hat{k}_H$	$\bar{d}$
	5 items	.211-.522	.205-.706	-.082	.077-.557	.106-.471	.035
	7 items	.267-.492	.269-.735	-.122	.119-.541	.157-.488	.012
	8 items	.338-.629	.397-.644	-.004	.043-.465	.078-.244	.073
	9 items	.225-.474	.306-.529	-.094	.255-.545	.233-.549	.03
	10 items	.468-.525	.724-.778	-.255	.559-.581	.559-.573	.004
	13 items*	E(k)=.354	$E(\hat{k}_H)$=.796	|d|=-.426	E(k)-.520	$E(\hat{k}_H)$=.622	d=-.102

*Represents single instance of test of this length.

As can be seen in Table 13, $\hat{k}_H$ provides a closer estimate of k at all test lengths for the heterogeneous tests. $\hat{k}_H$ has a more restricted range for the heterogeneous tests and the mean values for $\hat{k}_H$ are lower for the heterogeneous tests. Interestingly the opposite is true for k - i.e., k has a greater range and generally higher mean values for the heterogeneous tests. The result is that the mean values for k and $\hat{k}_H$ are closer. Additionally this is the first analysis for which $\hat{k}_H$ is not consistently higher than k.

Table 14 presents the absolute difference scores between the mean values of each coefficient for the homogeneous and the heterogeneous tests. For the most part k is higher for the homogeneous tests. $\hat{k}_H$, with the exception of grade 8 multiplication, is higher for the homogeneous tests. The coefficients of agreement on the other hand are generally higher for the heterogeneous tests.

TABLE 14

Absolute difference scores between mean coefficients from homogeneous tests and heterogeneous tests for all the reliability coefficients

Grade	Number Of Items	kappa	kappa	$p_{c(s)}$ (simple)	$p_{c(s)}$ (compound)
		Homo.-Heter.	Homo.-Heter.	Homo.-Heter.	Homo.-Heter.
3	5	-.035	.329	.085	-.028
	7	-.10	.293	.109	-.013
	10	-.091	.205	.066	.017
	13	-.166	.174	.04	-.013
5 (mult.)	5	-.128	.165	.073	.003
	7	-.11	.184	.026	.022
	10	-.056	.165	.054	.025
5 (divis.)	5	.143	.330	.016	-.129
	7	.127	.357	.09	-.099
	8	.089	.400	.033	-.108
6 (mult.)	5	.023	.039	-.126	-.075
	7	.008	.022	-.058	-.094
	9	.057	.028	-.095	-.052
6 (divis.)	5	.309	.273	-.17	-.111
	7	.248	.307	-.089	-.119
	8	.261	.361	-.114	-.061
7 (mult.)	5	.082	.017	-.21	-.025
	7	.097	.072	-.137	-.047
	9	-.03	.073	-.12	-.004
7 (divis.)	5	.094	.347	-.226	-.114
	7	.244	.410	-.175	-.126
	8	.289	.476	-.184	-.062
8 (mult.)	5	-.194	-.002	-.011	-.008
	7	-.136	-.008	-.016	.007
	9	-.263	-.02	-.004	.003
8 (divis.)	5	.348	.422	-.138	-.230
	7	.310	.443	-.034	-.208
	8	.586	.500	-.112	-.167
6-8 Combined (mult.)	5	.255	.137	-.175	-.069
	7	.214	.206	-.127	-.098
	9	.109	.255	-.125	-.052
6-8 Combined (divis.)	5	.270	.388	-.179	-.208
	7	.241	.383	-.105	-.218
	8	.356	.426	-.132	-.135

VI. Validation

The validation analysis consisted of computing Pearson Product Moment coefficients between the scores on the ITBS mathematics section and scores on the ISM mastery tests as given by MCPS for students in Grades 5 and 7. The math scores on the ITBS were considered an outside criterion of proficiency for these students. The validation analysis was restricted to examining the relationship between the ISM tests and the ITBS. Thus only predictive validity was looked at. It would have been possible using other methods to look at other types of validity. However, the availability of the ITBS scores made predictive validity appropriate to look at. Each student in both Grades 5 and 7 took a mastery test in multiplication and a mastery test in division, but the mastery tests were different for the different grades.

The mastery test in multiplication which the students in Grade 5 took was Multiplication 07-K and consisted of five problems in multiplication of 3-digit numbers by 2- and 3-digit numbers. The correlation between scores (i.e., number right) on the mastery test and scores on the ITBS mathematics section, was .343. The k coefficient computed for these 5 items was .397 (Table 15). The validity coefficient and the k coefficient are very similar in size and neither one is particularly high. The KR21 coefficient for these items is .731, much higher than the k or the validity coefficient. In norm-referenced test theory, validity is bounded by the square root of reliability. The square root of the kappa reliability coefficient is .630 and the square root of KR21 is .854. Thus for these items, the validity coefficient is clearly lower than the square root of the

TABLE 15

Validity coefficients and kappa coefficients for 4 mastery tests

Grade	Objective	Test Length	kappa Coefficient	Validity Coefficient	KR21
5	Multiplication 07-K	5 items	.397	.343	.731
7	Multiplication 08-L	5 items	.216	.119	.385
5	Division 08-J	5 items	.450	.287	.719
7	Division 10-N	5 items	.214	.438	.519

reliability coefficients. For these items then the results suggest that validity could be viewed as being bounded by the square root of k, as it is in the case of norm-referenced tests.

To continue with items measuring the same skill, multiplication, the students in Grade 7 took a mastery test MU 08-L which consisted of five items in multiplication of 3- and 4-digit numbers by 2-digit numbers. The correlation between scores on these items and scores on the ITBS mathematics section was .119. The k coefficient computed on these same items was .216. The KR21 coefficient for these items is .385. Both the validity coefficient and the correlation coefficient between the two sets of scores are very low. The square root of the k coefficient is .465 the square root of the KR21 coefficient is .620 and therefore, for these items as for the multiplication items for the fifth grade, the validity coefficient is smaller than the square root of the reliability. Again the results suggest that one could view the validity coefficient as being bounded by the square root of the reliability.

The division items which the fifth grade students took consisted of 5 problems in division by 1-digit numbers. The correlation between their scores on these items and the ITBS mathematics scores was .287. The k coefficient computed for these same items was .450, somewhat larger than the validity coefficient. The KR21 coefficient is .719. The square root of the k coefficient is .670 and the square root of the KR21 is .847. For these items the validity coefficient is bounded by the square root of reliability. The validity coefficient is very low suggesting little relationship between the division items and the ITBS items.

In the seventh grade, the students took a division mastery test DI 10-N which consisted of five items of division by 2-digit numbers. The correlation between these test scores and the ITBS math section scores was .438. The k coefficient computed on these same items was .214, somewhat lower than the validity coefficient. The KR21 coefficient was .519. The square root of the k coefficient was .462 and the square root of KR21 was .720. The validity coefficient for these items is much closer to the square root of the reliability coefficient yet it is still smaller.

VII. Relationship Between Norm-Referenced and Criterion-Referenced Coefficients

In order to investigate the relationship between norm-referenced and criterion-referenced coefficients, two analyses were performed. The first analysis consisted of computing KR21 coefficients for different test lengths and comparing them to the criterion-referenced coefficients for the same test lengths. The second analysis consisted of computing test-retest (TRT) coefficients on the mastery tests as given by MCPS (5-item tests for all objectives except MU 05-H, which was 10-items) and a random sample of items to construct tests of different lengths and then comparing the coefficients to the criterion-referenced coefficients for the same items. The items investigated were the same for the KR21 and the TRT analyses to permit comparisons between these coefficients. These analyses were performed for the three cut-off scores of 60, 80, and 100 percent. Each type of norm-referenced coefficient and its relationship to the criterion-referenced coefficients will be discussed separately.

A. Kuder-Richardson

Table 16 presents the mean values of the Kuder-Richardson coefficients, the TRT coefficients and the criterion-referenced coefficients for the cut-off scores of 60, 80, and 100 percent by test length. Grade level analyses are in the Appendix, Tables 69-71.

The lowest mean value for the KR coefficient is for the 5-item test, and the highest mean value is for the 13-item test. However, the mean values do not increase steadily as the test is lengthened. Rather the mean values of the KR coefficient fluctuate in the interval between 5 items and 13 items. At all cut-off scores the mean values of the KR coefficient are higher than the mean values of the k or $\hat{K}_H$ coefficients, but they are generally lower than the mean values of $p_{c(s)}$ (simple) and compound).

TABLE 16

Mean values of KR, TRT, and criterion-referenced coefficients by test length at 3 cut-off scores

Cut-Off Score	Test Length	$\bar{X}$ K-R	$\bar{X}$ TRT	E(k)	$E(\hat{k}_H)$	$E(p_{c(s)}(s))$	$E(p_{c(s)}(c))$
60%	5 items	.566	.368	.395	.443	.825	.941
	7 items	.889	.365	.361	.418	.861	.967
	8 items	.730	.509	.487	.351	.817	.938
	9 items	.620	.378	.292	.405	.832	.982
	10 items	.909	.636	.606	.777	.910	.977
	13 items	.931	.795	.703	.809	.927	.995
	Average	.774	.509	.474	.544	.862	.967
80%	5 items	.566	.368	.399	.438	.693	.890
	7 items	.889	.365	.384	.498	.690	.882
	8 items	.730	.509	.476	.553	.747	.884
	9 items	.620	.378	.358	.429	.724	.956
	10 items	.909	.636	.506	.774	.865	.961
	13 items	.931	.795	.354	.796	.802	.954
	Average	.774	.509	.413	.581	.754	.921
100%	5 items	.566	.368	.236	.439	.684	.756
	7 items	.889	.365	.225	.439	.728	.766
	8 items	.730	.509	.228	.458	.795	.769
	9 items	.620	.378	.333	.354	.729	.749
	10 items	.909	.636	.142	.696	.793	.878
	13 items	.931	.795	.264	.716	.806	.892
	Average	.774	.509	.238	.517	.388	.802

Table 17 presents the PPM correlations between the K-R coefficients and the mean values of the criterion-referenced coefficients for the different test lengths. For each test length, the mean values of each of the criterion-referenced coefficients were correlated by means of a Pearson Product Moment correlations, with the KR21 coefficients for the same items.

TABLE 17

PPM correlations between the KR21 and each criterion-referenced coefficient by test length and cut-off score

Cut-Off Score	Test Length	PPM between KR21 and k	PPM between KR21 and $\hat{k}_H$	PPM between KR21 and $p_{c(s)}(s)$	PPM between KR21 and $p_{c(s)}(c)$
60%	5 items	.566	.753**	-.396	-.103
	7 items	-.213	.969**	-.698*	-.550
	8 items	-.524	.995***	-.930*	-.812
	9 items	-.518	.977*	-.660	-.984*
80%	5 items	.745**	.637*	.432	-.044
	7 items	.654*	.902***	.824**	-.076
	8 items	.792	1.00***	.823	-.691
	9 items	.475	.934	-.134	-.684
100%	5 items	.135	.475	.420	.085
	7 items	-.758**	.946***	.831**	.916***
	8 items	.030	.981**	.980**	.992***
	9 items	-.898	.865	.756	.821

* $p < .05$
** $p < .01$
*** $p < .001$

The analysis of the relationship between internal consistency as measured by the KR21 coefficients and the criterion-referenced coefficients demonstrates evidence of some association between $\hat{k}_H$ and KR21 values and scattered, inconsistent associations between the KR21 values and other criterion-referenced coefficients. With the exception of $\hat{k}_H$, then, there is no general pattern of association between the KR21 coefficients and the criterion-referenced coefficients.

B. Test-Retest Coefficients

The mean values of the test-retest coefficients for the different test lengths appear in Table 16. Grade level analyses are in the Appendix, Tables 72-81. The mean values of the test-retest coefficients appear to be similar to the mean values of the k and $\hat{k}_H$ coefficients, but they are generally quite a lot smaller than the $p_{c(s)}$ coefficients.

Table 18 presents the PPM correlations between the test-retest coefficients and the criterion-referenced coefficients for the different test lengths by cut-off score. As was the case with the comparison of the KR21 coefficients and the criterion-referenced coefficients, a Pearson Product Moment coefficient was computed between the mean values of the criterion-referenced coefficients and the test-retest coefficients for each test length within each cut-off score.

TABLE 18

PPM correlations between the TRT and each criterion-referenced coefficient by test length and cut-off score

Cut-Off Score	Test Length	PPM between TRT and k	PPM between TRT and $\hat{k}_H$	PPM between TRT and $p_{c(s)}(s)$	PPM between TRT and $p_{c(s)}(c)$
60%	5 items	.739**	.693*	-.727*	-.497
	7 items	.630*	.265	-.281	-.356
	8 items	-.390	-.095	-.222	-.237
	9 items	.349	-.909	.790	+.994**
80%	5 items	.761**	.706*	.859***	.501
	7 items	.249	.082	.007	-.209
	8 items	.043	-.102	-.742	.136
	9 items	-.677	-.904	.319	.890
100%	5 items	-.174	.606*	.833**	.131
	7 items	.063	.144	.405	.120
	8 items	.950*	-.015	-.299	-.550
	9 items	.811	-.741	-.870	-.690

*$p < .05$
**$p < .01$
***$p < .001$

The analysis of the relationship between the TRT and the criterion-referenced coefficients reveals no pattern between any of the criterion-referenced coefficients and the test-retest coefficients. What relationships do exist are for the shorter test lengths.

VIII. Relationship Between k and the other Criterion-Referenced Coefficients

Another examination of the relationship between k and the other criterion-referenced coefficients was performed by computing Pearson Product Moment correlations between k and the other coefficients by cut-off score and test length. These results are presented in Table 19.

Overall, the highest degree of relationship is between k and $\hat{k}_H$ at the cut-off scores of 60 and 80 percent. Five of the eight possible correlations at these two cut-off scores are significant and positive. The 100 percent cut-off score has resulted in low values for k in previous analyses and the lack of correlation between k and $\hat{k}_H$ at this cut-off as well as the significant negative correlation at the 9-item test length reflects this. There is only one positive significant relationship between k and $p_{c(s)}(s)$ - 7-item tests with an 80 percent cut-off score - and the only significant relationships between k and $p_{c(s)}(c)$ are negative. It thus appears as if k and $\hat{k}_H$ are generally relatively close in their reliability determinations while k and $p_{c(s)}(s)$ and $p_{c(s)}(c)$ are very different in their assessment. Previous analyses have shown that the values of k and $\hat{k}_H$ were not terribly different from each other for the shorter tests while the values of $p_{c(s)}(c)$ in previous analyses have consistently been

much greater than the values of k. Therefore, the results of this correlational analysis are not surprising.

TABLE 19

PPM correlations between k and k, $\hat{k}_H$ and $p_{c(s)}(s)$, k and $p_{c(s)}(c)$ by test length and cut-off score across grade level

Cut-off Score	Test Length	r between k and $\hat{k}$	r between k and $p_{c(s)}(s)$	r between k and $p_{c(s)}(c)$
60%	5 items	.825**	-.414	-.624*
	7 items	.777**	-.377	-.522
	8 items	.574	-.287	.026
	9 items	-.659	-.298	.358
80%	5 items	.799**	.537	-.393
	7 items	.772**	.778**	-.101
	8 items	.949*	.462	-.769
	9 items	.321	-.918	-.917
100%	5 items	.199	.251	.454
	7 items	-.296	-.487	-.264
	8 items	.079	-.365	-.155
	9 items	-.958*	-.419	-.970*

*p < .05
**p < .01
***p < .001

Chapter 6

SUMMARY AND DISCUSSION

One of the problems inherent in the increased use of criterion-referenced tests for making decisions about educational progress has been the problem of determining the reliability of the tests. The determination of reliability for criterion-referenced tests has been conceptualized in several different ways. One involves the reliability of mastery classification decisions. A second concerns the reliability of criterion-referenced test scores and a third concerns the reliability of domain score estimates. While these three concepts of reliability are theoretically different from each other they all were developed as a result of the perception that norm-referenced reliability coefficients did not and could not provide appropriate information about criterion-referenced tests.

Since much of the use of criterion-referenced tests has been for the purposes of deciding on promotion in schools either from grade to grade or into another level or type of subject matter, the reliability of mastery classification decisions has been the focus of several conceptualizations of reliability for criterion-referenced tests. In subscribing to this view, one is concerned with the reliability of the decision that a student is a master or a nonmaster of material such that if a student were tested a second time would the mastery decision remain the same. The difficulty of course is that the use of criterion-referenced tests in instances of this type of decision-making often precludes the administration of the criterion-referenced test a second time.

In focusing on the view of reliability for criterion-referenced tests as the reliability of mastery classification decisions, three models for determining reliability have been proposed in the past several years: coefficient kappa, k, (Cohen, 1960) suggested for use with criterion-referenced tests by Swaminathan, Hambleton, and Algina, 1974, an estimate of coefficient kappa, $\hat{k}_H$ (Huynh, 1976) and the coefficient of agreement, $p_{c(s)}$ (Subkoviak, 1976). Coefficient k is based on two administrations of a criterion-referenced test while the estimate of kappa and the coefficient of agreement were developed for situations where only data from one test administration were available. When data are available from two test administrations one can get an empirical measure of reliability, kappa, but the availability of data from only one administration of a test is a common occurrence. Therefore the purpose of this research was to examine these three models for reliability in a situation where data were available from two test administrations so that the two estimates could be compared to a standard. The research involved the computation of the three coefficients for varying conditions which were manipulated. The variables which were examined for their effect on the reliability coefficients were: test length, cut-off score, ability level of the students, sample size, and heterogeneity of test content. Additionally, the following issues were also addressed: validity of the criterion-referenced tests, the relationship of norm-referenced reliability coefficients to the criterion-referenced coefficients, and the relationship of the standard, kappa, to the estimate of k, $\hat{k}_H$, based on a single test administration. The findings from each analysis will be summarized and discussed separately.

I. Test Length

H_1: All three reliability estimates , k, $\hat{k}_H$ and $p_{c(s)}$ will increase as the test are lengthened.

The results of the test length analysis showed that the predicted results were only partially attained. There was an increase in the mean values of all the reliability coefficients up to a test length of 8 items. Beyond that the results were unpredictable and there was no clear trend. Coefficient k was lowest for the longer test lengths and an increase in the other coefficients was difficult to detect due to the size of the standard errors. Thus, the results did not support the hypothesis about the effect of test length on the reliability coefficients.

The reason for the longer tests having the lower mean values for reliability may be that these tests were taken by the younger students. An examination of the data on these items indicates that for the 13-item test taken by the third graders, 23 percent of the students' scores changed from mastery to nonmastery or from nonmastery to mastery between the two test administrations. Those items with the most inconsistent performance were the last three items on the test, suggesting that fatigue and/or boredom might have been responsible for the inconsistent performance.

H_2: $\hat{k}_H$ will provide a good estimate of k for all test lengths.

The results of the difference score analysis indicated that $\hat{k}_H$ is only a good estimate of k for the shorter test lengths. Additionally $\hat{k}_H$ was consistently higher than k, particularly at the longer test

lengths. If fatigue and/or boredom were responsible for the inconsistent performance on the longer tests resulting in a low value for m, the empirical measure k computed on two test administrations, this would not affect an estimate of k, $\hat{k}_H$ based on only one test administration.

Some of the results of this test length analysis are similar to the results of other research reported in the literature. The findings about the size of the coefficients and their standard errors substantiate Subkoviak's (1978) findings, with one exception. Subkoviak (1978) reported that the Swaminathan et al. (1974) coefficient, k, produced relatively large errors of estimation for classroom size samples which occurred in this study. While the samples at each grade level are close to the size of two classrooms, this may not be a great enough increase over the size of one classroom to reduce the errors of estimation.

With respect to $\hat{k}_H$ Subkoviak (1978) reported that Huynh's (1976) method produced underestimates for all criterion levels for a 10-item test. This did not occur in this research as $\hat{k}_H$ was generally consistently higher than the standard, k. Subkoviak (1978) did find that this method resulted in low standard errors for classroom size samples, which also occurred in this study.

Finally, Subkoviak (1978) found that estimates of $p_{c(s)}$ tended to be overestimates of parameter values when the cut-off score was in the tails of the distribution, such as 80 percent, which is the cut-off score used for the test length analysis.

In summary then the results of the test length analysis in this research revealed that the coefficients increased with test length only up to a test length of 8 items. Additionally, there was little agreement among the coefficients, as measured by rank order correlation coefficients.

The question can be raised about which coefficient to use with a fixed mastery level. If there are data from two administrations, one would be able to compute k and get an empirical measure of the reliability. The difficulty is that k has been shown to have a very large standard error over .07 at least for sample sizes of approximately two classrooms. For larger sample sizes such as that for Grades 6-8 combined which was 191, the size of the standard errors is smaller.

In the case of one test administration, the $\hat{k}_H$ coefficient appears to be a reasonable estimate of k. The size of the $p_{c(s)}$ coefficient, particularly from the compound binomial model which was the recommended model, is so much greater than the standard k, as to suggest it might be an inflated estimate of reliability.

II. The Combined Effect of Cut-Off Score and Test Length

No formal hypothesis was proposed for this analysis because it was unclear how the change in cut-off score would affect the estimates.

The three reliability coefficients behaved differently as the cut-off score was raised. The three cut-off scores examined were 60, 80, and 100 percent.

With respect to $\hat{k}_H$, most of the maximum mean values occurred at the 80 percent cut-off score and there was a considerable decrease at the 100 percent cut-off score. The coefficients of agreement and k, however, were at their maximum mean value at the 60 percent cut-off score and the mean values declined steadily to the 100 percent cut-off score. Thus for $\hat{k}_H$, the results took the form of an inverted u and the results for k and for $p_{c(s)}$ (simple and compound) were a decreasing straight line.

Interestingly all the coefficients tend to be lowest when the cut-off score is set at 100 percent. There is a large difference in size between the $p_{c(s)}$ coefficients and the k and $\hat{k}_H$ coefficients. Yet each coefficient is lowest at the 100 percent cut-off. This result would suggest that setting a cut-off score so high would result in very low reliability.

With regard to how these results compare to other results reported in the literature, the results for $\hat{k}_H$ would support those of Huynh (1976) who found that the value of $\hat{k}_H$ increases as the cut-off score increases, up to a point and then it begins to decline. When the cut-off score is too low or too high (such as 100 percent) p_c is near one and this may explain why $\hat{k}_H$ declines after reaching a maximum value.

However, the results for the $p_{c(s)}$ coefficients are different from those reported by Subkoviak (1978) who reported that the procedure for computing the coefficient of agreement seemed to produce slight underestimates when the cut-off score was in the center of the test score distribution such as 50 percent and slight overestimates when the

cut-off score was in the tails of the distribution such as 80 percent. In this research the highest values of $p_{c(s)}$ occurred when the cut-off score was toward the center of the distribution (e.g., 60 percent) and the values decreased as the cut-off score was raised.

The only strongly positive relationships among the coefficients by test length occur at the 60 percent cut-off score. The correlation between k and $\hat{k}_H$ is 1.0; the correlation between k and $p_{c(s)}$ (simple) is .60 and the lowest, between k and $p_{c(s)}$ (compound) is .26. There is virtually no agreement between the coefficients at either the 80 or 100 percent cut-off score, despite the fact that all the coefficients reach their lowest mean values at the 100 percent cut-off score.

The use of $\hat{k}_H$ as an estimate of k appears reasonable only for the shorter test lengths using cut-off scores of 60 and 80 percent. Using a cut-off score of 100 percent, however, $\hat{k}_H$ is a good estimate of k only for the 9-item tests. However, even when $\hat{k}_H$ appears close in value to k, it is always consistently higher than k which may indicate that it overestimates k at all cut-off scores.

In summary, the analysis of the combined effect of cut-off score and test length revealed that both k and $p_{c(s)}$ (simple and compound) are highest for a cut-off score of 60 percent and lowest for a cut-off score of 100 percent. $\hat{k}_H$ is highest for a cut-off score of 80 percent and lowest for a cut-off score of 100 percent. There is some agreement among the coefficients at the 60 percent cut-off score, but none at the 80 or 100 percent cut-off scores. $\hat{k}_H$ provides a good estimate of k only for the shorter test lengths at the 60 and 80 percent cut-off scores.

III. The Combined Effect of Ability Level and Cut-Off Score

H_3: For homogeneous groups of high ability students or homogeneous low ability students, the change in cut-off score will have little effect on any of the three coefficients.

The hypothesis about the combined effect of ability level and cut-off score was supported for k, $\hat{k}_H$ and the $p_{c(s)}$ coefficients in that there was little consistent change in the size of the coefficients as the cut-off score was raised for either of the ability groups. Consistent with the analysis of test length and cut-off score the coefficients are at their lowest at the 100 percent cut-off score for both ability groups. Also consistent with previous analyses is the large standard errors for k.

With respect to differences in the sizes of the coefficients for the two ability groups, both k and $\hat{k}_H$ are slightly larger for the low ability group. An examination of the data reveals that the low ability group had more consistent decisions - consistent nonmastery than the high ability group which would result in higher values for k.

Neither of the $p_{c(s)}$ coefficients shows any consistency in being higher for one ability group or the other. Rank order correlation coefficients computed between k and the $p_{c(s)}$ coefficients in the two ability groups shows no relationship between the two coefficients for the low ability group while the correlations between $p_{c(s)}$ (simple) and k are positive at the 60 percent and 80 percent cut-off scores for the high ability group. The low ability group is more heterogeneous with respect to math achievement scores on the ITBS with scores which

range from the 2nd percentile to the 49th percentile. The high ability group, however, has scores which cluster between the 75th and 98th percentiles. The difference in the homogeneity of the groups may be the difference in the relationships between the $p_{c(s)}$ coefficients and k.

With respect to other research relative to the effect of ability level on the estimate of the $p_{c(s)}$ coefficients, Subkoviak (1978) stated that if a group is composed entirely or almost entirely of high ability students that fact alone would assure a large proportion of consistent mastery/mastery outcomes, and thus a high value for $p_{c(s)}$. The same should hold true for a group of low ability students - with a large proportion of consistent nonmastery/nonmastery outcomes $p_{c(s)}$ should also be high. This may explain why there is a difference within one objective according to cut-off score between the occurrence of higher mean values for the low ability or the high ability group. In this research neither of the $p_{c(s)}$ coefficients was consistently higher for the low or the high ability group. Rather for each cut-off score of 60, 80 or 100 percent the group with the highest proportion of consistent decisions has the highest mean values for $p_{c(s)}$. Thus, an examination of the data reveals that higher mean values for $p_{c(s)}$ for the high ability group reflect a high proportion of students mastering both tests. Similarly higher mean $p_{c(s)}$ values for the low ability group at the 100 percent cut-off is indicative of a high proportion of students <u>not</u> mastering either test.

H_4: $\hat{k}_H$ should provide a good estimate of k for homogeneous ability groups.

Two analyses investigated the relationship between k and $\hat{k}_H$. The rank order correlation coefficients computed between k and $\hat{k}_H$ differ by ability group. For the low ability group, there is a positive correlation between k and $\hat{k}_H$ only at the 100 percent cut-off score. At both the 60 percent and 80 percent cut-off scores there is no agreement between k and $\hat{k}_H$. An examination of the data reveals that the low ability group is quite heterogeneous in terms of math achievement on the ITBS, with percentile scores ranging from 02 to 49, while the percentile scores of the high ability group cluster in the 75-98 percentile range.

The positive correlation between k and $\hat{k}_H$ at the 100 percent cut-off score is perhaps indicative of low reliability for this cut-off score, regardless of the homogeneity or heterogeneity of the groups. In the analysis of the data on cut-off score all the reliability coefficients were lowest at the 100 percent cut-off and the positive relationship between k and $\hat{k}_H$ is possibly a reflection of that phenomenon.

There is a different relationship between k and $\hat{k}_H$ for the high ability group. At the 60 and 80 percent cut-off scores the correlation between k and $\hat{k}_H$ is positive--.70 at the 60 percent cut-off score and 1.0 at the 80 percent cut-off score. However, at the 100 percent cut-off score for this group the correlation between k and $\hat{k}_H$ is -.9. It is not totally clear why the correlation between k and $\hat{k}_H$ would be so strongly negative for this group, at this cut-off score.

The strong positive correlations between k and $\hat{k}_H$ at the 60 and 80 percent cut-off scores may be indicative of the high ability group being more homogeneous than the low ability group and thus resulting in $\hat{k}_H$ being a better estimate of k for this group. It is suspected that the 100 percent cut-off score results in such inconsistent performance that even for a high ability group where one would not expect k to be so low, $\hat{k}_H$ is simply a poor estimate of k.

An analysis of the mean difference scores between k and $\hat{k}_H$ resulted in $\hat{k}_H$ appearing to be a better estimate of k for the high ability group, at least at the 60 percent cut-off score. There is not much difference between $\hat{k}_H$ as an estimate of k for either ability groups for the 80 percent cut-off score and $\hat{k}_H$ is a poor estimate of k for both ability groups at the 100 percent cut-off score. $\hat{k}_H$'s being a better estimate of k for the high ability group may be a result of the greater homogeneity of the high ability group.

Interestingly, Huynh (1976) reports that test score variability has a positive relationship to $\hat{k}_H$. This is supported by this research in that the low ability group for whom the mean values of $\hat{k}_H$ were higher, has been found to be more heterogeneous in math achievement than the high ability group. The heterogeneity of ability then seems to increase the value of $\hat{k}_H$ yet it results in $\hat{k}_H$ being a poor estimate of k.

IV. Sample Size

H_5: The size of the sample will affect only the standard errors of all three reliability coefficients.

No hypothesis was proposed about the effect of the size of the sample on the reliability coefficients themselves.

With respect to k, the values of k show no predictable behavior from one sample size to another for the different objectives for which data were analyzed. Overall the mean values of k are slightly higher for the two classroom samples ($n \tilde{=} 50$). The size of the standard errors clearly reflects the differences in sample size as the standard errors were always larger for the smaller samples. This result substantiates Subkoviak's (1978) finding reported in the test length section that k produced rather large errors of estimation for classroom size samples. It should be noted, however, that even though the standard errors are smaller in the two classroom sample, they are still large (.072-.089) and quite a bit larger than the standard errors for $\hat{k}_H$ and the $p_{c(s)}$ coefficients.

The $\hat{k}_H$ coefficients are somewhat larger in the two classroom sample and, as predicted, the size of the standard errors is smaller in the two classroom sample. Thus, the hypothesis about $\hat{k}_H$ was supported. However, the standard errors for both one ($n \tilde{=} 25$) and two classroom ($n \tilde{=} 50$) samples across all grade levels, ranged from .022 to .110 for one classroom samples and from .022 to .093 for the two classroom samples.

This result is different from that which Subkoviak (1978) found. Subkoviak reported that Huynh's (1976) procedure resulted in estimates

with relatively small standard errors for classroom size samples. Since Subkoviak's (1978) finding was based on data from the SATs it is possible that the items from the ISM tests are more unreliable and thus the standard errors are larger in estimating the reliability.

Overall the analysis of effect of sample size on the $p_{c(s)}$ (simple) and $p_{c(s)}$ (compound) coefficients seems to indicate that the reliability coefficients for the one classroom samples were higher than those for the two classroom samples, at least at the 60 and 80 percent cut-off scores. When the seventh grade one classroom samples were compared to the total group Grades 6-8 combined taking the objective, the reliability coefficients were higher for the one classroom sample probably due to performance consistency not found in the more heterogeneous group. The prediction about the effect of sample size on this estimate estimate was that the effect would be on the standard errors rather than on the estimates. The size of the standard errors of both the $p_{c(s)}$ coefficients changes very little from one classroom to two classrooms. The standard errors are very small (.010 and below) for both groups.

While no effect from sample size was predicted on the coefficients, it should be noted that there were some changes in the rank order correlations between k and the other coefficients between the one and two classroom size samples. The changes, however, occurred only at the 60 percent cut-off score. There were strong positive correlations between k and $\hat{k}_H$ for both the one and two classroom samples at the 60 percent cut-off score, while there was no relationship between them at any other cut-off score in either sample size. In the two classroom

sample there was a moderate correlation, .60, between k and $p_{c(s)}$ (s) also at the 60 percent cut-off, but there was no agreement between k and $p_{c(s)}$ (s) or for that matter between k and $p_{c(s)}$ (c) at any other cut-off score for either sample size.

For the shorter test lengths with a cut-off score of 60 percent $\hat{k}_H$ appears to be a fairly good estimate of k for the two classroom sample although the estimates are somewhat higher than k. At the 80 percent and 100 percent cut-off scores the mean difference scores between the two coefficients are large for both sample sizes. Additionally $\hat{k}_H$ is consistently higher than k for all test lengths at these two cut-off scores.

In summary, the hypotheses about the effect of sample size on the standard errors of the reliability coefficients were supported for k and $\hat{k}_H$. For the $p_{c(s)}$ coefficients there was virtually no change in the standard errors for the different sample sizes. In general, $\hat{k}_H$ proved to be an overestimate of k for both sample sizes and was close in value to k only for short tests with a 60 percent cut-off score.

V. Item Heterogeneity

H_6: For heterogeneous test content both $\hat{k}_H$ and $p_{c(s)}$ will be lower than for homogeneous test content.

The effect of item heterogeneity was expected to be apparent only for coefficients $\hat{k}_H$ and $p_{c(s)}$ since the inclusion of different types of items would violate the assumptions of a binomial distribution. No effect on k was predicted for the heterogeneous test items.

While no effect on k was predicted for heterogeneous tests, the necessity for having tests or at least groups of items measuring only

one skill has been recommended in the literature on criterion-referenced testing for some time. Thus, one might consider that the determination of reliability for a heterogeneous test might lead to lower estimates of reliability. The results of the analysis of the effect of heterogeneous test items on k revealed that for the multiplication items for several groups the mean values for k were higher for the heterogeneous items than for the homogeneous items.

What is interesting is that regardless of level of the students, k does not appear to be sensitive to the inclusion of items measuring different skills on one test, at least with respect to the multiplication items. Perhaps the items which are purportedly measuring different multiplication skills are in fact not measuring different skills, resulting in higher mean values for k for supposedly heterogeneous items.

The results of the analyses of k for the division items (see Appendix, Tables 52-62) are closer to what had been predicted for k. For all the samples which took homogeneous and heterogeneous tests of division items, the mean values of k were lower for the heterogeneous items. The division items in the pool cover a wider range of objective levels than the multiplication items did and therefore it is more likely that more of the students had not mastered the skills measured by some of the items. Thus for the division items, construction of tests with items measuring different skills seems to have resulted in lower mean values for k.

The results of the analyses of $\hat{k}_H$ for heterogeneous items are very interesting and they do conform to the predictions about the

effect of heterogeneity of test content on $\hat{k}_H$. Since the estimate of $\hat{k}_H$ is based on the assumption that the distribution of the test score is binomial, (one assumption of which is that the probability of a correct response is constant for every item), it had been predicted that construction of a test with items which vary in content would have a substantial effect on $\hat{k}_H$. This is in fact the case since in every analysis the mean values of $\hat{k}_H$ were a great deal lower for the heterogeneous items than for the homogeneous items. While Subkoviak (1978) and Gross and Shulman (1978) have reported that the beta-binomial model appears to be quite robust with respect to violations of item homogeneity, this does not appear to be true in all instances in this research.

It had been predicted that the violation of test homogeneity would have a substantial effect on the coefficient of agreement for the same reason that it affected $\hat{k}_H$. The heterogeneous content of the tests did seem to affect $p_{c(s)}(s)$ and $p_{c(s)}(c)$, but in the opposite way than expected. Rather than decreasing the size of the coefficients in most cases the violation of test homogeneity increased the size of the coefficients. While the heterogeneity of test content seemed to decrease the precision of estimation for $p_{c(s)}(s)$, it appeared to increase the precision of estimation for $p_{c(s)}(c)$.

Subkoviak (1978) reported biased estimates for short tests and all these tests lengths are clearly short, but this does not explain why the $p_{c(s)}(s)$ and $p_{c(s)}(c)$ coefficients would increase for tests composed of heterogeneous items.

The rank order correlation between $p_{c(s)}(s)$ and k for the heterogeneous tests is -.58 and the correlation between k and $p_{c(s)}(c)$ is -.89. Thus, there is no agreement between k and the $p_{c(s)}$ coefficients for heterogeneous items. There also was no agreement between the $p_{c(s)}$ coefficients and k for tests of homogeneous items, using this cut-off score.

H_7: $\hat{k}_H$ will provide a poor estimate of k for heterogeneous test content.

Since the inclusion of heterogeneous items does violate the assumptions of a binomial model, one would expect $\hat{k}_H$ to be a poor estimate of k for these tests. However, the rank order correlation computed between k and $\hat{k}_H$ for the heterogeneous items was .944. This can be contrasted with the .315 correlation between k and $\hat{k}_H$ when the tests were homogeneous.

The stronger correlational relationship between k and $\hat{k}_H$ which exists for the heterogeneous items is substantiated by the mean difference scores between k and $\hat{k}_H$ for these items. For the heterogeneous items, $\hat{k}_H$ is a much better estimate of k than it is for the homogeneous items and this is the only analysis for which $\hat{k}_H$ is not consistently higher than k. Thus, the hypothesis about $\hat{k}_H$ being a poor estimate of k for the heterogeneous items was not supported. In fact $\hat{k}_H$ is a better estimate of k for these items than for any other analysis.

One aspect of the tests examined in this research which must be considered, particularly in light of the results of the analyses with heterogeneous items, is the construction of the tests. It is by no

means certain that the items from the different objectives constituted a particularly heterogeneous group. While item homogeneity has been touted as "one of the major benefits to result from operationally defining a content domain" (Berk, 1978), it is not clear whether much consideration was given to item homogeneity in constructing these tests. It is probable that no index of item homogeneity was utilized, but items were submitted to a judgmental process about their homogeneity. If the test items were less homogeneous by objective or less heterogeneous in the mixing of objectives, it is possible that the particular combinations of items selected in the random selection of items were responsible for the inconsistent results not only among the coefficients but also within the analyses for each individual coefficient. While a random sample of 100 combinations should be adequate to reduce bias due to selection of items, another explanation for these erratic results does not seem apparent.

VI. Validation

For both groups of students, both the reliability and the validity coefficients are very low (all below .50). This would suggest that the multiplication items taken by the students may not be particularly good measures of the skills they are intended to measure. Neither do the items reveal high performance consistency nor do they reveal a high degree of relationship to an outside criterion of performance. Thus, the validity was low. It does, however, seem to be the case for the criterion-referenced tests that validity is bounded by the square root of reliability as is the case for norm-referenced tests.

Overall the criterion-referenced test items in the two multiplication and two division mastery tests given by MCPS and analyzed here do not appear to have much validity in terms of an outside criterion of proficiency. The only group of items which had an even moderate relationship to the ITBS was the mastery test Division 10-N which the seventh graders took, and this relationship was not that strong ($r = .438$).

VII. Relationship Between Norm-Referenced and Criterion-Referenced Coefficients

H_8: The KR21 coefficients will have little relationship to any of the three criterion-referenced reliability estimates.

The analysis of the relationship between norm-referenced and criterion-referenced coefficients consisted of examining the relationship between the criterion-referenced coefficients and the KR21 coefficient as well as the relationship between the test-retest coefficients and test-retest coefficients. These relationships were examined by means of computing Pearson Product Moment correlation coefficients between the KR21 coefficients and the criterion-referenced coefficients and PPM correlations between the test-retest and criterion-referenced coefficients.

With regard to the KR21 coefficients, the only relationship which is consistent across cut-off scores is between the KR21 and the $\hat{k}_H$ coefficients. The correlations are positive for each test length and cut-off score and are significant in 9 out of 12 cases.

The relationship between the KR21 and $p_{c(s)}(s)$ coefficients is erratic, inverse for all test lengths at the 60 percent cut-off score, mixed at the 80 percent cut-off score and positive for all test lengths at the 100 percent cut-off score. The relationship between the KR21 and $p_{c(s)}(c)$ coefficients is also erratic, inverse for all test lengths at the 60 and 80 percent cut-off scores and highly positive at the 100 percent cut-off scores.

The relationship between the KR21 coefficients and k is also erratic, positive, and inverse at the same cut-off score, depending on test length.

These results seem to indicate that the KR21 coefficient and the criterion-referenced coefficients are assessing different types of reliability. The KR21 coefficients are a measure of the internal consistency of the tests while the criterion-referenced coefficients assess the consistency of mastery over two testings. Thus, although the two types of coefficients may be highly related they cannot be used interchangeably.

To some extent this analysis was limited by the small number of tests available for the analysis. However, the lack of a pattern of association between the KR21 and the criterion-referenced coefficients does seem to indicate that these norm-referenced and criterion-referenced indices are not assessing the same type of reliability.

H_9: The test-retest coefficients will show a moderate to strong relationship to all three criterion-referenced reliability coefficients.

The analysis of the relationship between the TRT and the criterion-referenced coefficients reveals no pattern between any of the two types of coefficients. Those relationships which do exist are for the shorter test lengths. Since the criterion-referenced indices are concerned with assessing consistency of mastery from one test administration to another (k) or the estimate thereof ($\hat{k}_H$, $p_{c(s)}(s)$, $p_{c(s)}(c)$), there would be expected a higher degree of relationship between them and the test-retest coefficients. Perhaps the limited number of available tests, particularly the 8- and 9-item tests, resulted in correlations which were not very meaningful.

The mastery level may be a crucial difference in that a change of one point from pre-test to post-test might not affect the TRT coefficient much while if the one point change were the difference between mastery and nonmastery, it would affect the criterion-referenced coefficient. Thus the relationship between the TRT and the criterion-referenced coefficients would not be very strong.

In conclusion, each of the three reliability coefficients examined has been found to have somewhat different properties although similarities do exist.

Coefficient k which can only be calculated when there are two test administrations generally has the lowest values and the largest standard errors. The size of k increases with increased test length, decreases when a cut-off score is set at an extreme point such as 100

percent, and seems to be maximized by larger sample sizes which also results in smaller standard errors. The heterogeneity of test content had a mixed effect on k, but it is suspected that truly heterogeneous test items might cause k to have lower values particularly if the item difficulties were different.

The estimate of kappa, $\hat{k}_H$, generally had slightly larger mean values than the mean values of k and smaller standard errors. The mean values of $\hat{k}_H$ increased with test length, and the mean values were highest for a cut-off score of 80 percent. The change in sample size did not predictably affect $\hat{k}_H$ and the mean values of $\hat{k}_H$ were highest for the low ability groups. The violation of test item homogeneity appeared to reduce $\hat{k}_H$.

The coefficient of agreement $p_{c(s)}$ from both the simple and the compound binomial models was the highest reliability coefficient with the smallest standard errors, across all analyses. Part of the reason for the high values of $p_{c(s)}$ is the difference in scaling between $p_{c(s)}$ and the kappa coefficients. The mean values of $p_{c(s)}$ did increase with test length and with respect to cut-off score, the maximum mean values occurred at the 60 percent cut-off score. The size of the $p_{c(s)}$ coefficient varied for ability level by cut-off score and test length, being higher for the high ability groups at the lower cut-off scores and higher for the low ability groups at the 100 percent cut-off score. The $p_{c(s)}$ coefficients' mean values were higher for one classroom samples than for the larger group, and the heterogeneity of test content increased the size of the mean values of $p_{c(s)}$.

The use of $\hat{k}_H$ as an estimate of k can be recommended only under certain conditions. The difference score analyses generally revealed that $\hat{k}_H$ overestimated k for all conditions except heterogeneous test content. $\hat{k}_H$ appears to be the best estimate of k for the following conditions: for short tests; for short tests with cut-off scores in the middle of the test score distribution (such as 60 percent); for homogeneous ability groups; for larger sample sizes with a cut-off score in the middle of the test score distribution; and for heterogeneous test content.

While with two test administrations one could compute the standard k, the size of the standard errors of k are so large that it is questionable how much information one is getting. $\hat{k}_H$ has been shown to consistently overestimate k somewhat, but the standard errors for $\hat{k}_H$ are very small. Thus, in some instances one might choose to administer a test only once and then calculate $\hat{k}_H$, knowing that it will usually be an overestimate, but that the standard errors are small. With the appropriate testing situation the use of $\hat{k}_H$ as an estimate could be recommended instead of scheduling a double test administration in order to calculate k.

Another possibility with respect to the apparent overestimate of k by $\hat{k}_H$ is that k may be an underestimate. There are certain performance factors which cannot be controlled and k, the standard may be too low due to these uncontrollable factors. If this were the case then $\hat{k}_H$ might be a more accurate estimate of reliability. $\hat{k}_H$ can now be viewed in a somewhat different light. This possibility would suggest that $\hat{k}_H$ might be regarded as an upper bound to reliability

and thus a low value for $\hat{k}_H$ should be regarded as a signal of questionable reliability for a test.

Thus the decision to use one of these coefficients must be made by considering the length and composition of the test, the cut-off score used for mastery, and the size of the sample. The intent of this research has been to provide information about the behavior of these three reliability coefficients based on actual criterion-referenced test data which were used to make decisions about students' progress in a mathematics program in a public school system. As such many of the findings should be generalizable to similar situations and practitioners should thus be provided with realistic expectations eliminating the need to extrapolate from findings generated through simulations or from nonrepresentative test data.

TABLE 20

TEST-RETEST DATA COLLECTION PROCEDURES

School	Grade	Number of Classes	Number of Examinees	Tests Administered: Day 1 Placement	Day 1 Mastery	Day 2 Placement	Day 2 Mastery
1	3	1	25	W-1	MU 05-H	W-1	MU 05-H
2	3	1	25	W-1	MU 05-H	W-1	MU 05-H
3	5	1	25	W-2	MU 07-K, DI 08-J	W-2	MU 07-K, DI 08-J
4	5	1	25	W-2	MU 07-K, DI 08-J	W-2	MU 07-K, DI 08-J
5	6	1	25	W-2	MU 08-L, DI 10-N	W-2	MU 08-L, DI 10-N
6	6	1	25	W-2	MU 08-L, DI 10-N	W-2	MU 08-L, DI 10-N
7	6	2	25	W-2	MU 08-L, DI 10-N	W-2	MU 08-L, DI 10-N
8	7	2	50	W-2	MU 08-L, DI 10-N	W-2	MU 08-L, DI 10-N
9	7	2	50	W-2	MU 08-L, DI 10-N	W-2	MU 08-L, DI 10-N
10	8	2	50	W-2	MU 08-L, DI 10-N	W-2	MU 08-L, DI 10-N
11	8	2	50	W-2	MU 08-L, DI 10-N	W-2	MU 08-L, DI 10-N

Placement Tests
W-1: Whole Numbers Test, Level 1, for students in Grades 3-4.
W-2: Whole Numbers Test, Level 2, for students in Grades 5-8.

Mastery Tests
MU 05-H: Multiplication 05-H, for 2nd semester, Grade 3.
MU 07-K: Multiplication 07-K, for 1st semester, Grade 5.
MU 08-L: Multiplication 08-L, for 2nd semester, Grade 5.
DI 08-J: Division 08-J, for 2nd semester, Grade 4.
DI 10-N: Division 10-N, for 2nd semester, Grade 6.

The grade levels for which these tests are considered appropriate, as indicated in the opposite column, refer to grade level placement within the Montgomery County (Md.) Public School (MCPS) curriculum.

Levels A, B=Kindergarten
Levels C, D=Grade 1
Levels E, F=Grade 2
Levels G, H=Grade 3
Levels I, J=Grade 4
Levels K, L=Grade 5
Levels M, N=Grade 6
Levels O, P=Grade 7
Levels Q, R=Grade 9

TABLE 21

MULTIPLICATION ITEMS FROM W-1 PLACEMENT TEST
AND FROM MU 05-H MASTERY TEST

These items were taken by the students in the 3rd grade sample.
(The letters and numbers on the sides refer to the particular objective which the items are assessing.)

MU04-H	23 x 2	201 x 4	312 x 3	429 x 2
MU05-H	68 x 2	876 x 3	709 x 4	290 x 5
MU06-I	75 x 6	860 x 8	497 x 7	847 x 9

MULTIPLICATION 05-H
FORM C

709 x 4	95 x 5	326 x 3	68 x 4	780 x 2
54 x 4	96 x 2	839 x 3	816 x 5	74 x 3

TABLE 22

MULTIPLICATION ITEMS FROM W-2 PLACEMENT TEST

These items were taken by the students in Grades 5-8 in the sample. (The letters and numbers on the sides refer to the particular objective which the items are assessing.)

PAGE 3 PLACEMENT W2-5/8

75	860	407	847	
x6	x 8	x 7	x 9	MU06-I
	705	800	354	
	x80	x100	x90	
67 x 10 = _____		510 x 10 = _____		MU07-K
809	80	7003	2867	
x39	x68	x78	x49	MU08-L
	23	201	312	420
MU04-H	x2	x 4	x 3	x 2
	68	876	709	290
MU05-H	x2	x 3	x 4	x 5

TABLE 23

DIVISION ITEMS FROM W-2 PLACEMENT TEST

These items were taken by the students in Grades 5-8 in the sample. (The letters and numbers on the sides refer to the particular objective which the items are assessing.)

Write any remainders after R in the quotient. example: 61R3

$3\overline{)63}$	$5\overline{)80}$	$2\overline{)79}$	$3\overline{)38}$	DI05-H
$3\overline{)872}$	$7\overline{)945}$	$2\overline{)419}$	$4\overline{)819}$	DI08-J
DI10-N	$44\overline{)4199}$		$86\overline{)4370}$	
	$12\overline{)7540}$		$28\overline{)8589}$	

TABLE 24

JOINT DISTRIBUTION OF SCORES ON TEST FORMS 1 and 2*

Form 1 (x) \ Form 2 (y)	0	1	2	3	4	5	6	7	8	9	10
0	0002	0006	0011	0013	0012	0008	0004	0002	0000	0000	0000
1	0006	0024	0050	0069	0068	0050	0028	0012	0004	0001	0000
2	0011	0050	0116	0174	0188	0152	0093	0043	0014	0003	0000
3	0013	0069	0174	0286	0338	0299	0201	0101	0036	0008	0000
4	0012	0068	0188	0338	0436	0421	0308	0169	0066	0017	0000
5	0008	0050	0152	0299	0421	0444	0354	0211	0090	0025	0000
6	0004	0028	0093	0201	0308	0354	0308	0200	0093	0028	0000
7	0002	0012	0043	0101	0169	0211	0200	0142	0072	0024	0000
8	0000	0004	0012	0036	0066	0090	0093	0072	0040	0014	0000
9	0000	0001	0003	0008	0017	0025	0028	0024	0014	0006	0000
10	0000	0000	0000	0001	0002	0003	0004	0004	0003	0001	0000

*Each entry in the body of Table 3 represents the proportion of examinees that would obtain score x on Form 1 and y on Form 2. Decimal points are omitted.

TABLE 25

TEST LENGTH
GRADE 3
OBJECTIVE: MULTIPLICATION 05-H

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
3	5 items	.522	.088	.706	.068	.807	.055	.940	.024
	7 items	.439	.077	.735	.045	.786	.043	.942	.023
	10 items	.468	.059	.778	.022	.819	.021	.969	.013
	13 items*	.354		.796		.802		.954	

TABLE 26

TEST LENGTH
GRADE 5
OBJECTIVE: MULTIPLICATION 07-K

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
5	5 items	.411	.109	.636	.042	.785	.032	.908	.022
	7 items	.431	.109	.672	.024	.794	.017	.918	.017
	10 items*	.525		.724		.832		.947	

*Maximum number of items for this objective.

TABLE 27

TEST LENGTH
GRADE 5
OBJECTIVE: DIVISION 08-J

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.$(\hat{k})$	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
5	5 items	.499	.078	.569	.035	.745	.025	.856	.036
	7 items	.492	.085	.598	.015	.780	.012	.884	.011
	8 items*	.554		.644		.790		.882	

*Maximum number of items for this objective.

TABLE 28

TEST LENGTH
GRADES 6, 7, 8 AND 6-8 COMBINED
OBJECTIVE: MULTIPLICATION 08-L

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
6	5 items	.357	.097	.326	.049	.627	.040	.899	.059
	7 items	.357	.103	.383	.029	.618	.016	.865	.042
	9 items*	.474		.435		.677		.927	
7	5 items	.280	.105	.205	.090	.716	.048	.971	.045
	7 items	.321	.077	.269	.065	.626	.026	.947	.045
	9 items	.225		.306		.742		.994	
8	5 items	.211	.136	.414	.110	.702	.072	.945	.034
	7 items	.317	.106	.480	.062	.679	.041	.924	.019
	9 items	.282		.529		.758		.954	
6-8 Combined	5 items	.317	.063	.339	.039	.669	.043	.927	.042
	7 items	.358	.049	.402	.022	.638	.019	.894	.027
	9 items	.362		.448		.718		.947	

*Maximum number of items for this objective.

TABLE 29

TEST LENGTH
GRADES 6, 7, 8 AND 6-8 COMBINED
OBJECTIVE: DIVISION 10-N

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
6	5 items	.418	.074	.469	.058	.674	.043	.885	.037
	7 items	.437	.058	.519	.028	.676	.021	.873	.021
	8 items*	.514		.554		.729		.938	
7	5 items	.272	.102	.311	.090	.677	.036	.882	.095
	7 items	.267	.059	.379	.041	.639	.026	.835	.035
	8 items	.338		.397		.742		.935	
8	5 items	.425	.099	.551	.027	.718	.019	.766	.036
	7 items	.429	.108	.583	.011	.754	.009	.786	.013
	8 items	.629		.625		.764		.832	
6-8 Combined	5 items	.409	.045	.494	.024	.689	.019	.789	.034
	7 items	.416	.029	.540	.011	.699	.009	.778	.014
	8 items	.522		.573		.747		.864	

*Maximum number of items for this objective.

TABLE 30

CUT-OFF SCORE
GRADE 3
OBJECTIVE: MULTIPLICATION 05-H

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_c)$ (simple)	$s.e.(p_c)$ (simple)	$E(p_c)$ (compound)	$s.e.(p_c)$ (compound)
60%	5 items	.723	.068	.703	.069	.882	.037	.935	.015
	7 items	.754	.051	.743	.045	.915	.022	.961	.014
	10 items	.742	.045	.783	.022	.924	.012	.983	.008
	13 items*	.703		.809		.927		.995	
80%	5 items	.522	.088	.706	.068	.807	.055	.940	.024
	7 items	.439	.077	.735	.045	.786	.043	.942	.023
	10 items	.468	.059	.778	.022	.819	.021	.969	.013
	13 items	.354		.796		.802		.954	
100%	5 items	.309	.115	.664	.075	.744	.029	.853	.042
	7 items	.291	.108	.685	.052	.758	.018	.854	.038
	10 items	.259	.073	.704	.028	.783	.009	.875	.027
	13 items	.264		.716		.806		.892	

*Maximum number of items for this objective.

TABLE 31

CUT-OFF SCORE
GRADE 5
OBJECTIVE: MULTIPLICATION 07-K

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.433	.071	.630	.043	.836	.028	.944	.021
60%	7 items	.432	.049	.678	.024	.865	.016	.953	.011
	10 items*	.459		.728		.882		.958	
	5 items	.411	.109	.636	.042	.785	.032	.908	.022
80%	7 items	.431	.109	.672	.024	.794	.017	.918	.017
	10 items	.525		.724		.832		.947	
	5 items	.259	.156	.589	.047	.731	.016	.849	.029
100%	7 items	.217	.122	.613	.028	.760	.011	.867	.021
	10 items	.132		.635		.792		.905	

*Maximum number of items for this objective.

TABLE 32

CUT-OFF SCORE
GRADE 5
OBJECTIVE: DIVISION 08-J

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_c)$ (simple)	$s.e.(p_c)$ (simple)	$E(p_c)$ (compound)	$s.e.(p_c)$ (compound)
60%	5 items	.626	.067	.585	.033	.763	.025	.884	.032
	7 items	.644	.044	.637	.013	.802	.011	.932	.021
	8 items*	.593		.657		.810		.934	
80%	5 items	.499	.078	.569	.035	.745	.025	.856	.036
	7 items	.492	.085	.598	.015	.780	.012	.884	.011
	8 items	.554		.644		.790		.882	
100%	5 items	.291	.173	.489	.041	.813	.024	.805	.024
	7 items	.096	.175	.506	.019	.859	.012	.826	.014
	8 items	-.089		.511		.877		.830	

*Maximum number of items for this objective.

TABLE 33

CUT-OFF SCORE
GRADE 6
OBJECTIVE: MULTIPLICATION 08-L

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_c)$ (simple)	$s.e.(p_c)$ (simple)	$E(p_c)$ (compound)	$s.e.(p_c)$ (compound)
60%	5 items	.339	.097	.286	.051	.811	.039	.989	.019
	7 items	.351	.117	.339	.031	.847	.020	.997	.006
	9 items*	.381		.419		.779		.974	
80%	5 items	.357	.097	.326	.049	.627	.040	.899	.059
	7 items	.357	.103	.383	.029	.618	.016	.865	.042
	9 items	.474		.435		.677		.927	
100%	5 items	.251	.128	.289	.052	.642	.033	.655	.061
	7 items	.286	.101	.315	.032	.725	.027	.695	.035
	9 items	.332		.331		.788		.740	

*Maximum number of items for this objective.

TABLE 34

CUT-OFF SCORE
GRADE 7
OBJECTIVE: MULTIPLICATION 08-L

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.184	.211	.149	.082	.919	.031	.999	.003
60%	7 items	.125	.193	.174	.063	.947	.018	1.00	0.0
	9 items*	.345		.260		.883		1.00	
	5 items	.280	.105	.205	.090	.716	.048	.971	.045
80%	7 items	.321	.077	.269	.065	.626	.026	.947	.045
	9 items	.225		.306		.742		.994	
	5 items	.332	.094	.205	.086	.548	.028	.719	.119
100%	7 items	.403	.076	.239	.058	.604	.027	.691	.069
	9 items	.464		.266		.668		.722	

*Maximum number of items for this objective.

TABLE 35

CUT-OFF SCORE
GRADE 8
OBJECTIVE: MULTIPLICATION 08-L

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.112	.122	.374	.117	.852	.037	.988	.018
60%	7 items	.109	.102	.434	.069	.881	.018	.993	.009
	9 items*	.119		.514		.845		.974	
	5 items	.211	.136	.414	.110	.702	.072	.945	.034
80%	7 items	.317	.106	.480	.062	.679	.041	.924	.019
	9 items	.258		.529		.758		.954	
	5 items	.251	.076	.387	.112	.636	.029	.779	.069
100%	7 items	.231	.081	.427	.067	.684	.018	.786	.031
	9 items	.175		.453		.726		.787	

*Maximum number of items for this objective.

TABLE 36

CUT-OFF SCORE
GRADES 6, 7 and 8 COMBINED
OBJECTIVE: MULTIPLICATION 08-L

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.$(\hat{k})$	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
60%	5 items	.273	.083	.290	.041	.852	.033	.993	.012
	7 items	.283	.085	.341	.024	.882	.017	.997	.004
	9 items*	.324		.426		.822		.980	
80%	5 items	.317	.063	.339	.039	.669	.043	.927	.042
	7 items	.358	.049	.402	.022	.638	.019	.894	.027
	9 items	.362		.448		.718		.947	
100%	5 items	.297	.042	.315	.042	.611	.024	.691	.064
	7 items	.333	.029	.347	.026	.679	.023	.717	.031
	9 items	.360		.367		.736		.749	

*Maximum number of items for this objective.

TABLE 37

CUT-OFF SCORE
GRADE 6
OBJECTIVE: DIVISION 10-N

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.479	.086	.449	.061	.792	.029	.959	.028
60%	7 items	.505	.047	.509	.029	.837	.010	.965	.005
	8 items*	.544		.544		.817		.966	
	5 items	.418	.074	.469	.058	.674	.043	.885	.037
80%	7 items	.437	.058	.519	.028	.676	.021	.873	.021
	8 items	.515		.554		.729		.938	
	5 items	.248	.074	.419	.063	.691	.020	.721	.020
100%	7 items	.310	.061	.447	.032	.753	.013	.747	.014
	8 items	.384		.457		.779		.757	

*Maximum number of items for this objective.

TABLE 38

CUT-OFF SCORE
GRADE 7
OBJECTIVE: DIVISION 10-N

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.285	.181	.260	.093	.868	.026	.976	.043
60%	7 items	.400	.109	.309	.049	.892	.015	.996	.012
	8 items*	.393		.359		.853		.995	
	5 items	.272	.102	.311	.090	.677	.036	.882	.095
80%	7 items	.266	.059	.379	.041	.639	.026	.835	.035
	8 items	.338		.397		.742		.935	
	5 items	.153	.084	.293	.086	.596	.028	.645	.077
100%	7 items	.138	.052	.329	.040	.664	.013	.663	.043
	8 items	.134		.342		.694		.674	

*Maximum number of items for this objective.

TABLE 39

CUT-OFF SCORE
GRADE 8
OBJECTIVE: DIVISION 10-N

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.452	.089	.556	.027	.759	.016	.833	.044
60%	7 items	.434	.044	.609	.010	.798	.007	.882	.025
	8 items*	.412		.632		.791		.875	
	5 items	.425	.099	.551	.027	.718	.019	.766	.036
80%	7 items	.429	.108	.583	.011	.754	.009	.786	.013
	8 items	.629		.625		.764		.832	
	5 items	.316	.176	.479	.029	.774	.013	.764	.018
100%	7 items	.279	.160	.497	.013	.828	.007	.809	.012
	8 items	.253		.503		.847		.875	

*Maximum number of items for this objective.

TABLE 40

CUT-OFF SCORE
GRADES 6, 7 and 8 COMBINED
OBJECTIVE: DIVISION 10-N

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.457	.052	.477	.026	.793	.013	.900	.029
60%	7 items	.484	.036	.533	.011	.832	.004	.929	.009
	8 items*	.495		.566		.814		.921	
	5 items	.409	.045	.494	.024	.689	.019	.789	.034
80%	7 items	.415	.029	.540	.011	.699	.009	.778	.014
	8 items	.522		.573		.747		.864	
	5 items	.257	.059	.443	.026	.693	.012	.689	.012
100%	7 items	.264	.039	.469	.012	.753	.007	.737	.006
	8 items	.281		.477		.776		.756	

*Maximum number of items for this objective.

TABLE 41

LOW ABILITY
GRADE 5
OBJECTIVE: MULTIPLICATION 07-K

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_c)$ (simple)	$s.e.(p_c)$ (simple)	$E(p_c)$ (compound)	$s.e.(p_c)$ (compound)
60%	5 items	.395	.109	.661	.043	.815	.030	.902	.022
	7 items	.387	.089	.704	.024	.844	.018	.937	.021
	10 items*	.420		.748		.857		.952	
80%	5 items	.376	.115	.654	.045	.783	.029	.885	.035
	7 items	.449	.108	.682	.027	.802	.018	.896	.023
	10 items	.527		.733		.836		.909	
100%	5 items	.349	.165	.593	.053	.777	.018	.839	.029
	7 items	.336	.139	.611	.033	.810	.015	.857	.019
	10 items	.257		.627		.842		.882	

*Maximum number of items for this objective.

TABLE 42

HIGH ABILITY
GRADE 5
OBJECTIVE: MULTIPLICATION 07-K

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_c)$ (simple)	$s.e.(p_c)$ (simple)	$E(p_c)$ (compound)	$s.e.(p_c)$ (compound)
60%	5 items	.389	.132	.516	.098	.849	.034	.986	.021
	7 items	.398	.067	.577	.055	.877	.020	.991	.012
	10 items*	.423		.646		.896		.972	
80%	5 items	.399	.153	.542	.087	.760	.054	.946	.028
	7 items	.371	.119	.598	.047	.757	.031	.943	.017
	10 items	.488		.658		.798		.963	
100%	5 items	.147	.168	.509	.086	.678	.037	.876	.045
	7 items	.067	.133	.547	.049	.712	.019	.881	.029
	10 items	-.026		.579		.756		.862	

*Maximum number of items for this objective.

TABLE 43

LOW ABILITY
GRADE 5
OBJECTIVE: DIVISION 08-J

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.668	.126	.671	.051	.816	.036	.844	.027
60%	7 items	.719	.059	.715	.020	.843	.022	.946	.009
	8 items*	.752		.728		.865		.970	
	5 items	.697	.131	.645	.056	.824	.039	.891	.041
80%	7 items	.688	.122	.665	.024	.849	.017	.906	.027
	8 items	.738		.709		.864		.878	
	5 items	.436	.297	.558	.067	.873	.023	.878	.035
100%	7 items	.189	.328	.571	.029	.900	.011	.926	.029
	8 items	-.040		.576		.910		.951	

*Maximum number of items for this objective.

TABLE 44

HIGH ABILITY
GRADE 5
OBJECTIVE: DIVISION 08-J

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.526	.112	.482	.049	.730	.043	.933	.042
60%	7 items	.525	.113	.539	.021	.777	.021	.945	.021
	8 items*	.401		.567		.775		.910	
	5 items	.341	.128	.482	.053	.682	.027	.829	.055
80%	7 items	.348	.096	.519	.026	.722	.009	.846	.029
	8 items	.419		.563		.734		.884	
	5 items	.165	.206	.409	.063	.767	.034	.754	.031
100%	7 items	.004	.183	.428	.033	.827	.021	.789	.014
	8 items	-.139		.433		.850		.803	

*Maximum number of items for this objective.

TABLE 45

LOW ABILITY
GRADES 6-8 COMBINED
OBJECTIVE: MULTIPLICATION 08-L

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_c)$ (simple)	$s.e.(p_c)$ (simple)	$E(p_c)$ (compound)	$s.e.(p_c)$ (compound)
60%	5 items	.249	.184	.428	.066	.814	.036	.989	.015
	7 items	.267	.177	.488	.037	.848	.017	.992	.009
	9 items*	.479		.556		.824		.967	
80%	5 items	.314	.109	.456	.061	.692	.047	.927	.029
	7 items	.346	.065	.512	.033	.689	.026	.913	.014
	9 items	.404		.563		.753		.938	
100%	5 items	.229	.153	.416	.063	.666	.029	.793	.048
	7 items	.189	.175	.449	.037	.721	.023	.797	.029
	9 items	.113		.469		.764		.808	

*Maximum number of items for this objective.

TABLE 46

HIGH ABILITY
GRADES 6-8 COMBINED
OBJECTIVE: MULTIPLICATION 08-L

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.245	.097	.224	.062	.879	.030	.996	.008
60%	7 items	.262	.097	.266	.040	.908	.016	.999	.002
	9 items*	.239		.354		.838		.990	
	5 items	.295	.083	.279	.063	.674	.044	.935	.043
80%	7 items	.324	.061	.345	.036	.619	.024	.888	.031
	9 items	.323		.386		.714		.953	
	5 items	.276	.042	.264	.062	.583	.025	.646	.077
100%	7 items	.318	.031	.298	.036	.653	.022	.667	.049
	9 items	.361		.318		.716		.696	

*Maximum number of items for this objective.

TABLE 47

LOW ABILITY
GRADES 6-8 COMBINED
OBJECTIVE: DIVISION 10-N

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.426	.133	.442	.058	.689	.031	.864	.081
60%	7 items	.449	.086	.502	.027	.727	.013	.914	.042
	8 items*	.394		.529		.722		.860	
	5 items	.362	.086	.435	.058	.667	.032	.737	.042
80%	7 items	.273	.101	.468	.028	.740	.011	.763	.019
	8 items	.518		.520		.720		.815	
	5 items	.118	.179	.352	.062	.803	.023	.752	.038
100%	7 items	-.012	.158	.365	.032	.868	.013	.815	.018
	8 items	-.086		.369		.889		.837	

*Maximum number of items for this objective.

TABLE 48

HIGH ABILITY
GRADES 6-8 COMBINED
OBJECTIVE: DIVISION 10-N

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.255	.065	.358	.027	.835	.014	.942	.021
60%	7 items	.257	.082	.411	.011	.871	.005	.962	.005
	8 items*	.304		.453		.842		.955	
	5 items	.289	.069	.398	.024	.671	.019	.793	.038
80%	7 items	.345	.033	.454	.009	.649	.009	.749	.015
	8 items	.413		.478		.734		.872	
	5 items	.215	.063	.366	.025	.630	.012	.624	.013
100%	7 items	.243	.051	.395	.011	.697	.008	.674	.008
	8 items	.275		.404		.726		.697	

*Maximum number of items for this objective.

TABLE 49

ONE CLASS
GRADE 3
OBJECTIVE: MULTIPLICATION 05-H

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_c)$ (simple)	$s.e.(p_c)$ (simple)	$E(p_c)$ (compound)	$s.e.(p_c)$ (compound)
60%	5 items	.659	.083	.736	.071	.851	.044	.885	.023
	7 items	.681	.068	.772	.044	.880	.028	.942	.019
	10 items	.657	.064	.807	.022	.889	.014	.968	.013
	13 items*	.606		.828		.889		.989	
80%	5 items	.458	.119	.722	.076	.817	.043	.927	.029
	7 items	.363	.103	.743	.051	.821	.027	.935	.024
	10 items	.404	.073	.788	.025	.832	.016	.965	.021
	13 items	.324		.799		.837		.937	
100%	5 items	.179	.167	.659	.092	.824	.024	.879	.041
	7 items	.098	.162	.674	.063	.848	.022	.892	.029
	10 items	-.008	.139	.688	.035	.879	.016	.900	.019
	13 items	-.114		.696		.905		.883	

*Maximum number of items for this objective.

TABLE 50

ONE CLASS
GRADE 5
OBJECTIVE: MULTIPLICATION 07-K

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_c)$ (simple)	$s.e.(p_c)$ (simple)	$E(p_c)$ (compound)	$s.e.(p_c)$ (compound)
	5 items	.498	.144	.588	.083	.886	.027	.974	.019
60%	7 items	.531	.113	.638	.049	.906	.016	.980	.009
	10 items*	.500		.696		.911		.990	
	5 items	.393	.180	.612	.075	.815	.051	.947	.028
80%	7 items	.432	.184	.659	.042	.805	.037	.939	.018
	10 items	.395		.709		.853		.959	
	5 items	.389	.177	.588	.075	.702	.045	.878	.052
100%	7 items	.387	.129	.619	.044	.715	.024	.891	.041
	10 items	.333		.647		.733		.934	

*Maximum number of items for this objective.

TABLE 51

ONE CLASS
GRADE 5
OBJECTIVE: DIVISION 08-J

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_c)$ (simple)	$s.e.(p_c)$ (simple)	$E(p_c)$ (compound)	$s.e.(p_c)$ (compound)
	5 items	.495	.129	.518	.080	.796	.043	.953	.025
60%	7 items	.493	.086	.575	.037	.838	.018	.964	.009
	8 items*	.559		.605		.824		.991	
	5 items	.473	.137	.529	.079	.698	.055	.903	.039
80%	7 items	.517	.137	.574	.037	.698	.025	.905	.022
	8 items	.600		.609		.755		.914	
	5 items	.252	.186	.476	.086	.713	.026	.763	.051
100%	7 items	.024	.192	.503	.043	.762	.019	.762	.026
	8 items	-.185		.512		.783		.758	

*Maximum number of items for this objective.

TABLE 52 .

ONE CLASS
GRADE 7
OBJECTIVE: MULTIPLICATION 08-L

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
	5 items	.107	.256	.135	.095	.945	.021	.999	.007
60%	7 items	.078	.202	.143	.074	.968	.012	1.00	0.0
	9 items*	.353		.218		.926		1.00	
	5 items	.287	.184	.192	.109	.767	.048	.976	.052
80%	7 items	.283	.084	.244	.089	.668	.038	.962	.033
	9 items	.358		.271		.800		.995	
	5 items	.261	.117	.204	.110	.542	.030	.728	.126
100%	7 items	.309	.108	.229	.086	.567	.021	.679	.088
	9 items	.355		.259		.616		.661	

*Maximum number of items for this objective.

TABLE 53

ONE CLASS
GRADE 7
OBJECTIVE: DIVISION 10-N

Cut-off Score	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_c)$ (simple)	$s.e.(p_c)$ (simple)	$E(p_c)$ (compound)	$s.e.(p_c)$ (compound)
	5 items	.105	.251	.249	.108	.914	.029	.996	.011
60%	7 items	.000		.291	.059	.932	.016	1.00	0.0
	8 items*	.000		.344		.901		1.00	
	5 items	.174	.186	.309	.105	.748	.037	.962	.044
80%	7 items	.206	.075	.384	.049	.688	.026	.923	.023
	8 items	.338		.392		.803		.988	
	5 items	.089	.105	.311	.098	.574	.037	.786	.062
100%	7 items	.085	.063	.354	.047	.616	.017	.779	.017
	8 items	.087		.371		.638		.780	

*Maximum number of items for this objective.

TABLE 54

HETEROGENEOUS ITEMS RANDOM SELECTION OF 100 COMBINATIONS
MULTIPLICATION ITEMS N = 52

MASTERY LEVEL = 80% (Number of Items in pool = 22)

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_c)$ (simple)	$s.e.(p_c)$ (simple)	$E(p_c)$ (compound)	$s.e.(p_c)$ (compound)
3	5 items	.557	.119	.377	.199	.722	.094	.968	.041
	7 items	.539	.116	.442	.179	.677	.085	.955	.036
	10 items	.559	.098	.573	.095	.753	.049	.952	.018
	13 items	.520	.081	.622	.052	.762	.024	.967	.014

TABLE 55

HETEROGENEOUS ITEMS RANDOM SELECTION OF 100 COMBINATIONS
MULTIPLICATION ITEMS N = 54

MASTERY LEVEL = 80% (Number of Items in pool = 26)

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	$s.e.(\hat{k})$	$E(p_c)$ (simple)	$s.e.(p_c)$ (simple)	$E(p_c)$ (compound)	$s.e.(p_c)$ (compound)
5	5 items	.539	.104	.471	.099	.712	.054	.905	.044
	7 items	.541	.116	.488	.068	.768	.048	.896	.032
	10 items	.581	.105	.559	.037	.778	.024	.922	.021

TABLE 56

HETEROGENEOUS ITEMS RANDOM SELECTION OF 100 COMBINATIONS
DIVISION ITEMS N = 54

MASTERY LEVEL = 80% (Number of Items in pool = 17)

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
5	5 items	.346	.229	.239	.151	.729	.138	.985	.036
	7 items	.365	.176	.241	.123	.690	.121	.983	.032
	8 items	.465	.166	.244	.153	.757	.134	.990	.018

TABLE 57

HETEROGENEOUS ITEMS RANDOM SELECTION OF 100 COMBINATIONS
MULTIPLICATION ITEMS N = 80

MASTERY LEVEL = 80% (Number of Items in pool = 17)

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
6	5 items	.334	.145	.287	.118	.753	.082	.974	.036
	7 items	.349	.081	.361	.087	.676	.060	.959	.032
	9 items	.417	.099	.407	.068	.772	.049	.979	.014

TABLE 58

HETEROGENEOUS ITEMS RANDOM SELECTION OF 100 COMBINATIONS
MULTIPLICATION ITEMS N = 57

MASTERY LEVEL = 80% (Number of Items in pool = 17)

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
7	5 items	.198	.207	.188	.098	.836	.078	.996	.019
	7 items	.224	.157	.197	.104	.763	.074	.994	.014
	9 items	.255	.212	.233	.097	.862	.056	.998	.008

TABLE 59

HETEROGENEOUS ITEMS RANDOM SELECTION OF 100 COMBINATIONS
MULTIPLICATION ITEMS N = 54

MASTERY LEVEL = 80% (Number of Items in pool = 17)

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
8	5 items	.405	.139	.416	.125	.713	.050	.953	.051
	7 items	.453	.115	.488	.072	.695	.043	.938	.028
	9 items	.545	.091	.549	.038	.762	.032	.951	.019

TABLE 60

HETEROGENEOUS ITEMS RANDOM SELECTION OF 100 COMBINATIONS
MULTIPLICATION ITEMS N = 191

MASTERY LEVEL = 80% (Number of Items in pool = 17)

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
6, 7, & 8 Combined	5 items	.364	.102	.333	.111	.756	.071	.969	.042
	7 items	.403	.057	.418	.068	.702	.048	.958	.027
	9 items	.473	.056	.471	.045	.785	.038	.973	.013

TABLE 61

HETEROGENEOUS ITEMS RANDOM SELECTION OF 100 COMBINATIONS
DIVISION ITEMS N = 80

MASTERY LEVEL = 80% (Number of Items in pool = 22)

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
6	5 items	.109	.159	.196	.126	.844	.128	.996	.010
	7 items	.189	.176	.212	.120	.765	.132	.992	.016
	8 items	.253	.141	.193	.110	.843	.150	.999	.003

TABLE 62

HETEROGENEOUS ITEMS RANDOM SELECTION OF 100 COMBINATIONS
DIVISION ITEMS N = 57

MASTERY LEVEL = 80% (Number of Items in pool = 22)

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
7	5 items	.178	.182	.122	.094	.900	.102	.999	.002
	7 items	.193	.198	.109	.088	.851	.115	.999	.006
	8 items	.225	.370	.078	.064	.913	.126	1.00	0.0

TABLE 63

HETEROGENEOUS ITEMS RANDOM SELECTION OF 100 COMBINATIONS
DIVISION ITEMS N = 54

MASTERY LEVEL = 80% (Number of Items in pool = 22)

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
8	5 items	.077	.159	.129	.103	.856	.107	.996	.012
	7 items	.119	.163	.140	.112	.788	.111	.994	.014
	8 items	.043	.108	.125	.101	.876	.124	.999	.005

TABLE 64

HETEROGENEOUS ITEMS RANDOM SELECTION OF 100 COMBINATIONS
DIVISION ITEMS N = 191

MASTERY LEVEL = 80%

(Number of Items in pool = 22)

Grade	Test Length	E(k)	s.e.(k)	$E(\hat{k})$	s.e.($\hat{k}$)	$E(p_c)$ (simple)	s.e.(p_c) (simple)	$E(p_c)$ (compound)	s.e.(p_c) (compound)
6, 7, & 8 Combined	5 items	.139	.152	.106	.092	.868	.109	.997	.012
	7 items	.175	.140	.157	.099	.804	.119	.996	.008
	8 items	.166	.184	.147	.091	.879	.127	.999	.003

TABLE 65

KUDER-RICHARDSON AND CRITERION-REFERENCED COEFFICIENTS FOR 5-ITEM TESTS WITH A CUT-OFF SCORE OF 60 PERCENT

Grade	Objective	K-R	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
3	MU05-H	.826	.722	.614	.842	.933
5	MU07-K	.731	.357	.572	.832	.944
5	DI08-J	.719	.556	.580	.759	.899
6	MU08-L	.635	.221	.360	.817	.969
6	DI10-N	.674	.379	.458	.770	.965
7	MU08-L	.385	.383	.207	.921	1.00
7	DI10-N	.519	.296	.197	.876	1.00
8	MU08-L	.265	.175	.499	.876	.980
8	DI10-N	.426	.524	.548	.747	.792
6-8 Combined	MU08-L	.486	.254	.374	.861	.987
6-8 Combined	DI10-N	.558	.458	.469	.779	.881

TABLE 66

KUDER-RICHARDSON AND CRITERION-REFERENCED COEFFICIENTS FOR 5-ITEM TESTS WITH A CUT-OFF SCORE OF 80 PERCENT

Grade	Objective	K-R	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
3	MU05-H	.826	.505	.755	.859	.923
5	MU07-K	.731	.450	.578	.742	.882
5	DI08-J	.719	.523	.546	.761	.818
6	MU08-L	.635	.387	.353	.613	.933
6	DI10-N	.674	.492	.499	.716	.875
7	MU08-L	.385	.160	.236	.676	.989
7	DI10-N	.519	.384	.336	.659	.811
8	MU08-L	.265	.055	.159	.564	1.00
8	DI10-N	.426	.555	.566	.715	.797
6-8 Combined	MU08-L	.486	.444	.285	.611	.967
6-8 Combined	DI10-N	.558	.441	.509	.705	.795

TABLE 67

KUDER-RICHARDSON AND CRITERION-REFERENCED COEFFICIENTS FOR 5-ITEM TESTS WITH A CUT-OFF SCORE OF 100 PERCENT

Grade	Objective	K-R	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
3	MU05-H	.826	.386	.718	.775	.884
5	MU07-K	.731	.412	.545	.699	.823
5	DI08-J	.719	.069	.484	.811	.786
6	MU08-L	.635	.352	.360	.642	.638
6	DI10-N	.674	.103	.412	.711	.735
7	MU08-L	.385	.232	.277	.555	.827
7	DI10-N	.519	.170	.236	.578	.581
8	MU08-L	.265	.332	.511	.651	.815
8	DI10-N	.426	.047	.464	.786	.774
6-8 Combined	MU08-L	.486	.332	.397	.617	.772
6-8 Combined	DI10-N	.558	.154	.428	.704	.686

TABLE 68

KUDER-RICHARDSON AND CRITERION-REFERENCED COEFFICIENTS FOR 7-ITEM TESTS WITH A CUT-OFF SCORE OF 60 PERCENT

Grade	Objective	K-R	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
3	MU05-H	.800	.807	.679	.902	.967
5	MU07-K	.887	.482	.676	.835	.949
5	DI08-J	.815	.626	.642	.801	.950
6	MU08-L	.634	.510	.368	.844	.997
6	DI10-N	.695	.481	.485	.830	.970
7	MU08-L	.342	-.024	.142	.959	1.00
7	DI10-N	.508	.397	.300	.900	1.00
8	MU08-L	.750	.133	.477	.891	1.00
8	DI10-N	.772	.435	.604	.790	.872
6-8 Combined	MU08-L	.592	.386	.366	.885	1.00
6-8 Combined	DI10-N	.703	.487	.521	.829	.934

TABLE 69

KUDER-RICHARDSON AND CRITERION-REFERENCED COEFFICIENTS FOR 7-ITEM TESTS WITH A CUT-OFF SCORE OF 80 PERCENT

Grade	Objective	K-R	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
3	MU05-H	.800	.362	.674	.719	.970
5	MU07-K	.887	.325	.661	.791	.873
5	DI08-J	.815	.598	.608	.783	.890
6	MU08-L	.634	.322	.408	.629	.858
6	DI10-N	.695	.536	.498	.667	.879
7	MU08-L	.342	.281	.245	.624	.955
7	DI10-N	.508	.234	.375	.638	.848
8	MU08-L	.750	.289	.519	.700	.949
8	DI10-N	.772	.503	.531	.697	.786
6-8 Combined	MU08-L	.592	.358	.425	.650	.909
6-8 Combined	DI10-N	.703	.416	.531	.697	.786

TABLE 70

KUDER-RICHARDSON AND CRITERION-REFERENCED COEFFICIENTS FOR 7-ITEM TESTS WITH A CUT-OFF SCORE OF 100 PERCENT

Grade	Objective	K-R	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
3	MU05-H	.800	.235	.616	.737	.867
5	MU07-K	.887	.199	.593	.787	.849
5	DI08-J	.815	.039	.518	.847	.833
6	MU08-L	.634	.241	.339	.725	.711
6	DI10-N	.695	.262	.424	.752	.743
7	MU08-L	.342	.459	.219	.588	.633
7	DI10-N	.508	.218	.327	.656	.660
8	MU08-L	.750	.113	.471	.670	.849
8	DI10-N	.772	.154	.493	.824	.805
6-8 Combined	MU08-L	.592	.298	.372	.672	.742
6-8 Combined	DI10-N	.703	.262	.459	.750	.733

TABLE 71

KUDER-RICHARDSON AND CRITERION-REFERENCED COEFFICIENTS FOR 8-, 9-, 10- AND 13-ITEM TESTS WITH A CUT-OFF SCORE OF 60 PERCENT

8-ITEM TESTS

Grade	Objective	K-R	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
5	DI08-J	.836	.593	.657	.810	.934
6	DI10-N	.712	.544	.544	.817	.966
7	DI10-N	.551	.393	.359	.853	.995
8	DI10-N	.820	.412	.631	.791	.875
6-8 Combined	DI10-N	.731	.495	.566	.814	.921
9-ITEM TESTS						
6	MU08-L	.671	.381	.419	.779	.974
7	MU08-L	.461	.345	.260	.883	1.00
8	MU08-L	.718	.119	.514	.845	.974
6-8 Combined	MU08-L	.631	.324	.425	.822	.980
10-ITEM TESTS						
3	MU05-H	.960	.753	.825	.938	.995
7	MU07-K	.858	.459	.728	.882	.958
13-ITEM TESTS						
3	MU05-H	.931	.703	.809	.927	.995

TABLE 72

KUDER-RICHARDSON AND CRITERION-REFERENCED COEFFICIENTS FOR 8-, 9-, 10- AND 13-ITEM TESTS WITH A CUT-OFF SCORE OF 80 PERCENT

8-ITEM TESTS

Grade	Objective	K-R	k	$\hat{k}$	$p_o^{(s)}$	$p_o^{(c)}$
5	DI08-J	.836	.533	.644	.790	.882
6	DI10-N	.712	.514	.554	.729	.938
7	DI10-N	.551	.338	.375	.638	.848
8	DI10-N	.820	.629	.625	.764	.832
6-8 Combined	DI10-N	.731	.368	.566	.814	.921
9-ITEM TESTS						
6	MU08-L	.671	.474	.435	.677	.927
7	MU08-L	.461	.239	.306	.742	.994
8	MU08-L	.718	.282	.529	.758	.954
6-8 Combined	MU08-L	.631	.435	.448	.718	.947
10-ITEM TESTS						
3	MU05-H	.960	.486	.819	.847	.964
7	MU07-K	.858	.525	.728	.882	.958
13-ITEM TESTS						
3	MU05-H	.931	.354	.796	.802	.954

TABLE 73

KUDER-RICHARDSON AND CRITERION-REFERENCED COEFFICIENTS FOR 8-, 9-, 10- AND 13-ITEM TESTS WITH A CUT-OFF SCORE OF 100 PERCENT

Grade	Objective	K-R	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
8-ITEM TESTS						
5	DI08-J	.836	-.089	.511	.877	.830
6	DI10-N	.712	.384	.456	.779	.757
7	DI10-N	.551	.134	.342	.694	.674
8	DI10-N	.820	.253	.502	.847	.826
6-8 Combined	DI10-N	.731	.281	.477	.776	.756
9-ITEM TESTS						
6	MU08-L	.671	.332	.331	.788	.740
7	MU08-L	.461	.464	.266	.668	.722
8	MU08-L	.718	.175	.453	.726	.787
6-8 Combined	MU08-L	.631	.360	.367	.736	.749
10-ITEM TESTS						
3	MU05-H	.960	.152	.756	.793	.850
7	MU07-K	.858	.132	.635	.792	.905
13-ITEM TESTS						
3	MU05-H	.931	.264	.716	.806	.892

TABLE 74

TEST-RETEST AND CRITERION-REFERENCED COEFFICIENTS FOR 5-ITEM TESTS WITH A CUT-OFF SCORE OF 60 PERCENT

Grade	Objective	TRT	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
3	MU05-H	.745	.722	.614	.842	.933
5	MU07-K	.467	.357	.572	.832	.944
5	DI08-J	.677	.556	.580	.759	.899
6	MU08-L	.433	.221	.360	.817	.969
6	DI10-N	.578	.379	.458	.770	.965
7	MU08-L	-.076	.383	.207	.921	1.00
7	DI10-N	.248	.296	.197	.876	1.00
8	MU08-L	.097	.175	.499	.876	.980
8	DI10-N	.448	.524	.548	.747	.792
6-8 Combined	MU08-L	.271	.254	.374	.861	.987
6-8 Combined	DI10-N	.512	.458	.469	.779	.881

TABLE 75

TEST-RETEST AND CRITERION-REFERENCED COEFFICIENTS FOR 5-ITEM TESTS WITH A CUT-OFF SCORE OF 80 PERCENT

Grade	Objective	TRT	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
3	MU05-H	.745	.505	.755	.859	.923
5	MU07-K	.467	.397	.587	.764	.883
5	DI08-J	.677	.450	.565	.731	.871
6	MU08-L	.433	.601	.397	.660	.850
6	DI10-N	.578	.492	.471	.661	.893
7	MU08-L	-.076	.216	.274	.733	.994
7	DI10-N	.248	.214	.253	.650	.934
8	MU08-L	.097	.235	.534	.778	.942
8	DI10-N	.448	.435	.541	.717	.742
6-8 Combined	MU08-L	.271	.444	.419	.713	.915
6-8 Combined	DI10-N	.512	.441	.484	.675	.754

TABLE 76

TEST-RETEST AND CRITERION-REFERENCED COEFFICIENTS FOR 5-ITEM TESTS WITH A CUT-OFF SCORE OF 100 PERCENT

Grade	Objective	TRT	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
3	MU05-H	.745	.386	.718	.775	.884
5	MU07-K	.467	.412	.545	.699	.823
5	DI08-J	.677	.069	.484	.811	.786
6	MU08-L	.433	.352	.360	.642	.638
6	DI10-N	.578	.103	.412	.711	.735
7	MU08-L	-.076	.232	.277	.555	.827
7	DI10-N	.248	.170	.236	.578	.581
8	MU08-L	.097	.332	.511	.651	.815
8	DI10-N	.448	.047	.464	.786	.774
6-8 Combined	MU08-L	.271	.332	.397	.617	.772
6-8 Combined	DI10-N	.512	.154	.428	.704	.686

TABLE 77

TEST-RETEST AND CRITERION-REFERENCED COEFFICIENTS FOR 7-ITEM TESTS WITH A CUT-OFF SCORE OF 60 PERCENT

Grade	Objective	TRT	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
3	MU05-H	.764	.807	.679	.902	.967
5	MU07-K	.258	.482	.676	.835	.949
5	DI08-J	.706	.626	.642	.801	.950
6	MU08-L	.645	.510	.368	.844	.997
6	DI10-N	.464	.481	.485	.830	.970
7	MU08-L	.489	-.024	.142	.959	1.00
7	DI10-N	.234	.397	.300	.900	1.00
8	MU08-L	-.039	.133	.477	.891	1.00
8	DI10-N	.591	.435	.604	.790	.872
6-8 Combined	MU08-L	.453	.386	.366	.885	1.00
6-8 Combined	DI10-N	.555	.487	.521	.829	.934

TABLE 78

TEST-RETEST AND CRITERION-REFERENCED COEFFICIENTS FOR 7-ITEM TESTS WITH A CUT-OFF SCORE OF 80 PERCENT

Grade	Objective	TRT	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
3	MU05-H	.764	.362	.674	.719	.970
5	MU07-K	.258	.325	.661	.791	.873
5	DI08-J	.706	.598	.608	.783	.890
6	MU08-L	.645	.322	.408	.629	.858
6	DI10-N	.464	.536	.498	.667	.879
7	MU08-L	.489	.281	.245	.624	.955
7	DI10-N	.234	.234	.375	.638	.848
8	MU08-L	-.039	.289	.519	.700	.949
8	DI10-N	.591	.503	.579	.761	.801
6-8 Combined	MU08-L	.453	.358	.425	.650	.909
6-8 Combined	DI10-N	.555	.416	.531	.697	.786

TABLE 79

TEST-RETEST AND CRITERION-REFERENCED COEFFICIENTS FOR 7-ITEM TESTS WITH A CUT-OFF SCORE OF 100 PERCENT

Grade	Objective	TRT	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
3	MU05-H	.764	.235	.616	.737	.867
5	MU07-K	.258	.199	.593	.787	.849
5	DI08-J	.706	.039	.518	.847	.833
6	MU08-L	.645	.241	.339	.725	.711
6	DI10-N	.464	.262	.424	.752	.743
7	MU08-L	.489	.459	.219	.588	.633
7	DI10-N	.234	.218	.327	.656	.660
8	MU08-L	-.039	.113	.471	.670	.849
8	DI10-N	.591	.154	.493	.824	.805
6-8 Combined	MU08-L	.453	.298	.372	.672	.742
6-8 Combined	DI10-N	.555	.262	.459	.750	.733

TABLE 80

TEST-RETEST AND CRITERION-REFERENCED COEFFICIENTS FOR 8-, 9-, 10- AND 13-ITEM TESTS WITH A CUT-OFF SCORE OF 60 PERCENT

8-ITEM TESTS

Grade	Objective	TRT	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
5	DI08-J	.201	.593	.657	.810	.934
6	DI10-N	.652	.544	.544	.817	.966
7	DI10-N	.437	.393	.359	.853	.995
8	DI10-N	.627	.412	.631	.791	.875
6-8 Combined	DI10-N	.629	.495	.566	.814	.921
9-ITEM TESTS						
6	MU08-L	.304	.381	.419	.779	.974
7	MU08-L	.519	.345	.260	.883	1.00
8	MU08-L	.312	.119	.514	.845	.974
6-8 Combined	MU08-L	.377	.324	.425	.822	.980
10-ITEM TESTS						
3	MU05-H	.851	.753	.825	.938	.995
7	MU07-K	.421	.459	.728	.882	.958
13-ITEM TESTS						
3	MU05-H	.795	.703	.809	.927	.995

TABLE 81

TEST-RETEST AND CRITERION-REFERENCED COEFFICIENTS FOR 8-, 9-, 10- AND 13-ITEM TESTS WITH A CUT-OFF SCORE OF 80 PERCENT

8-ITEM TESTS

Grade	Objective	TRT	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
5	DI08-J	.201	.533	.644	.790	.882
6	DI10-N	.652	.514	.554	.729	.938
7	DI10-N	.437	.338	.375	.638	.848
8	DI10-N	.627	.629	.625	.764	.832
6-8 Combined	DI10-N	.629	.368	.566	.814	.921
9-ITEM TESTS						
6	MU08-L	.304	.474	.435	.677	.927
7	MU08-L	.519	.239	.306	.742	.994
8	MU08-L	.312	.282	.529	.758	.954
6-8 Combined	MU08-L	.377	.435	.448	.718	.947
10-ITEM TESTS						
3	MU05-H	.851	.552	.819	.847	.964
7	MU07-K	.421	.525	.728	.882	.958
13-ITEM TESTS						
3	MU05-H	.795	.354	.796	.802	.954

TABLE 82

TEST-RETEST AND CRITERION-REFERENCED COEFFICIENTS FOR 8-, 9-, 10- AND 13-ITEM TESTS WITH A CUT-OFF SCORE OF 100 PERCENT

Grade	Objective	TRT	k	$\hat{k}$	$p_o(s)$	$p_o(c)$
8-ITEM TESTS						
5	DI08-J	.201	-.089	.511	.877	.830
6	DI10-N	.652	.384	.456	.779	.757
7	DI10-N	.437	.134	.342	.694	.674
8	DI10-N	.627	.253	.502	.847	.826
6-8 Combined	DI10-N	.629	.281	.477	.776	.756
9-ITEM TESTS						
6	MU08-L	.304	.332	.331	.788	.740
7	MU08-L	.519	.464	.266	.668	.722
8	MU08-L	.312	.175	.453	.726	.787
6-8 Combined	MU08-L	.377	.360	.367	.736	.749
10-ITEM TESTS						
3	MU05-H	.851	.152	.756	.793	.850
7	MU07-K	.421	.132	.635	.792	.905
13-ITEM TESTS						
3	MU05-H	.795	.264	.716	.806	.892

References

Algina, J. and Noe, M. J. A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. Journal of Educational Measurement, 1978, 15 (2), 101-110.

Berk, R. A. Criterion-referenced test item analysis and validation. Paper presented at the first annual Johns Hopkins University, National Symposium on Educational Research, Washington, D. C., October 1978.

Block, J. H., Criterion-referenced measurement: Potential, School Review, 1971, 79, 289-297.

Block, J. H., ed., Mastery Learning: Theory and Practice, New York: Holt, Rinehart, and Winston, Inc., 1971.

Bloom, Benjamin S., "Learning for mastery," UCLA-CSEIP, Evaluation Comment, 1, No. 2 (1968).

Bloom, Benjamin S., et. al. (eds.), Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain, New York: David McKay Co., Inc., 1956.

Brennan, R. L. and Kane, M. T., An index of dependability for mastery tests, Journal of Educational Measurement, 1977, 14 (3), 277-289.

Carroll, John B., "A model of school learning," Teachers College Record, 64 (1963), 723-33.

Carver, R. P., Special problems in measuring change with psychometric devices, in Evaluative Research Strategies and Methods, Washington: American Institutes for Research, 1970.

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

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

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Ebel, R. L., Criterion-referenced measurement: Limitations, School Review, 1971, 79, 282-287.

Ebel, R. L., The case for norm-referenced measurement, Educational Researcher, 1978, 7(11), pp. 3-5.

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

Glaser, R., A criterion-referenced test in W. J. Popham (Ed.) Criterion-referenced Measurement: An Introduction, Educational Technology Publications, Inc., Englewood Cliffs, N.J., 1971.

Gross, A. and Shulman, V. The Applicability of the Beta Binomial Model for Criterion-Referenced Testing.

Hambleton, R. K., Decision-making in instructional programs, Review of Educational Research, 1974, 44 (4), 371-400.

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

Hambleton, R. K., Swaminathan, H., Algina, J., and Coulson, D. B., Criterion-referenced testing and measurement: A review of technical issues and development, Review of Educational Research, Winter 1978, 48, 1-47.

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

Huynh, H., On the reliability of decisions in domain-referenced testing, Journal of Educational Measurement, 1976, 13, 253-264.

Huynh, H., Reliability of mastery classifications. Psychometrika, 1978, 43 (3), 317-325.

Livingston, S. A., Criterion-referenced applications of classical test theory, Journal of Educational Measurement, 1972, 9, 31. (b).

Logsdon, D. M., A study of the meaningfulness of criterion- and norm-referenced reliability indices in assessing the validity of mastery rates. Unpublished doctoral dissertation, Florida State University, 1979.

Lord, F. M. and Novick, M. R., Statistical Theories of Mental Test Scores, Reading, Mass.: Addison-Wesley, 1968.

Meskauskas, J. A., Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting, Review of Educational Research, 1976, 46 (1), 133-158.

Millman, J., Criterion-referenced measurement in W. J. Popham (Ed.) Evaluation in Education: Current Applications, Berkeley, California, McCutchan Publishing Co., 1974.

Popham, W. J. (Ed.), Criterion-referenced measurement: An introduction, Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Popham, W. J., Educational Evaluation, Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Popham, W. J., The case for criterion-referenced measurements, Educational Research, 1978, 7 (11), 6-10.

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

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

Subkoviak, M., Estimating reliability from a single administration of a criterion-referenced test, Journal of Educational Measurement, 1976, 13, 265-275.

Subkoviak, M., Empirical investigation of procedures for estimating reliability for mastery tests, Journal of Educational Measurement, 1978, 15, 111-116.

Subkoviak, M., The reliability of mastery classification decisions, Paper presented at the first annual Johns Hopkins University National Symposium on Educational Research, Washington, D. C., October 1978.

Swaminathan, H., Hambleton, R. K., and Algina, J., Reliability of criterion-referenced tests: A decision-theoretic formulation, Journal of Educational Measurement, 1974, 11, 263-268.

Washburne, Carleton W., Educational measurements as a key to individualizing instruction and promotions, Journal of Educational Research, 1922, 5, 195-206.

Winkler, R. L., An Introduction to Bayesian Inference and Decision, New York: Halt, Rinehart, and Winston, Inc., 1972.