8-1976

A Comparison of the Kuder Richardson Formula 20 and Kappa as Estimates of the Reliability of Criterion-Referenced Tests

Judith E. Moyer
Western Michigan University

A Comparison of the Kuder Richardson Formula 20 and Kappa as Estimates of the Reliability of Criterion-Referenced Tests

by

Judith E. Moyer

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

Western Michigan University
Kalamazoo, Michigan
August, 1976

ACKNOWLEDGEMENTS

This study is the culmination of the opportunity and responsibility awarded to me by Western Michigan University in the form of an E.P.D.A. grant in August, 1973. The grants in that series gave women and minority men the resources to complete one full year of study toward the doctoral degree. I thank the University and the U.S. Government for their support.

Thank you to my mentors, Dr. Mary Anne Bunda, Dr. Uldis Smidchens, and Dr. Ernest Stech for the challenges and encouragement they have provided me. Thank you to Jim Moyer for being perfectly wonderful.

I wish to express my gratitude to the Highland Park public school teachers who, between 1946 and 1959, did so much to enable me to be where I am now. I also acknowledge with appreciation the impact so many others have had on my development; especially, I hope that my child will feel that she has had as much freedom to "be" from her parents as I had from my mother and father.

Finally, thanks to Cathy Crawford for the evenings and Saturday's she spent typing this manuscript.

Judith E. Moyer

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

The Problem

Reliability of measurement is of primary importance in any scientific endeavor, and educational testing is no exception. Since the early 1900's, researchers have been developing theory and methods for estimating reliability. For any kind of measure, be it a scale, a ruler or a test, reliability has been accepted to mean consistency. Stanley (1971), however, points out that operationalizing the concept of reliability or consistency is not simply a process of selecting and using an available formula:

> There is no single, universal and absolute reliability coefficient for a test. Determination of reliability is as much a logical as a statistical problem. The appropriate allocation of variance from different sources calls for practical judgement of what use is to be made of the resulting statistical value (p.363).

This issue has become particularly clear with the wide scale use of criterion-referenced tests. Some knowledgeable researchers question the use of traditional methods of estimating reliability with criterion-referenced tests. They believe that the conditions which must be logically satisfied in order to use the traditional methods cannot be satisfied. As will be discussed in more detail later, the lack of variation among the scores on criterion-referenced tests suggests that traditional reliability estimates which largely depend on variation are inappropriate. However, this

position is not accepted by all researchers. Various educational researchers have maintained that traditional methods are appropriate with criterion-referenced tests (Ebel, 1975; Livingston, 1972), while others have maintained that the nature and use of criterion-referenced tests demand reinterpretation of traditional methods (Hambleton and Novick, 1973) or new methods altogether (Swaminathan, Hambleton and Algina, 1974). Clearly, there is a need for empirical data to clarify the nature of these theoretical positions.

The test-retest method of estimating reliability, wherein examinees are administered a test at one time and are subsequently administered the same test at another time, would seem to be the most logical and definitive means of operationalizing the concept of reliability. If individuals received the same scores on both tests, one could assume the measure to be reliable. This would be analogous to measuring a board with a ruler two times. If the results were the same both times, the carpenter could consider the measure to be a reliable one. However, students of human behavior have a problem in establishing the reliability of their measuring devices by this means that is not shared by all scientific researchers: that is, the human often changes as a result of being measured, so that subsequent measurement may be different, not necessarily because of errors in the measuring instrument, but because of changes that have occurred over time in the individuals being measured. For this reason and other practical considerations, such as time and money, other methods of estimating

test-retest reliability have been developed and are frequently used. Two commonly reported methods are parallel forms and KR-20.

Parallel forms, although they are used by commercial test publishers, suffer from the extreme difficulty of their development. Parallel forms must meet the following criteria: they must have equal mean item difficulties, equal mean item variances, equal variances of item variances, equal means of item covariances within and between tests as has been well explained by Horst (1966). The application of the theory, as might be expected, suffers in comparison to the prescription (Horst, 1966; Bohrnstedt, 1970). However, Bohrnstedt maintains that "roughly parallel" forms can be developed and Stanley (1971) says that preparing "parallel forms should not present undue difficulty" (p.405). Apparently the question to be answered is "what should these parallel forms uniformly do?" If the purpose of the parallel test is to rank each individual in the same order relative to others, then this should be possible to accomplish, even if the theoretical criteria are impossible to satisfy in their entirety.

KR-20, the second estimate of reliability, measures the internal consistency or homogeneity of the items of a test by treating each item as a parallel test with every other item. Horst (1966) points out that homogeneity, although it contributes to the estimation of reliability, is not, in and of itself, a measure of

reliability, because it does not "indicate solely the extent to which the measures yielded by the individual items can be relied upon" (p.262).

Obviously, these methods of estimating the reliability of traditional, norm-referenced tests are subject to some operational problems, but the introduction of criterion-referenced tests has heightened the educational researchers' awareness of the old problems as well as directed attention to possible new problems. Norm-referenced tests are expected to produce variability, to spread examinees along a continuum according to each individual's level of the quality being measured; therefore, most reliability measures depend on the variability of the scores. Criterion-referenced tests, however, are designed to measure whether or not examinees have attained a specific skill. Usually, they are short tests, administered soon after a period of instruction to determine which students have attained a skill and which have not. For this reason, variability among students is expected to be low, possibly nonexistent, if all the students have mastered the skill. Theoretically, then, as Stanley (1971) points out, "a criterion-referenced test can give reliable information even though its classically defined reliability coefficient equals zero" (p.435). In other words, if one could use the test-retest method in this real situation, the measure would be found to be reliable. In this case, the operationalization of the concept of consistency with a classical reliability coefficient would be, logically, a very poor

choice. Again, it should be remembered that since criterion-referenced tests are usually short, the test-retest method of determining reliability would also be a poor choice in that the memory effect, the phenomenon of examinees, on the retest, remembering their previous answers could produce spuriously high reliability coefficients.

Since the purpose of criterion-referenced tests is the classification of students into mastery and nonmastery categories on a specific skill, Millman (1974), Hambleton and Novick (1973), and Swaminathan, et. al. (1974) have recommended that the consistency of decisions regarding these two categories be the proper operationalization of the concept of reliability. The consistency of decisions across two tests is measured by kappa (κ), and these researchers urge that this consistency of the decision process, as measured by κ, is a reflection of the quality of the content and the use of the test for making mastery-nonmastery decisions.

In practice, little is known about κ, but it appears to provide an appropriate operational definition of reliability. The coefficient of agreement indicates the proportion of agreement about masters and nonmasters beyond what would have been expected by chance across two (or more) versions of a test.

The problem that this research will attempt to resolve is the following: do the KR-20 and kappa coefficients provide different estimates of the test-retest reliability of a series of pairs of tests which were designed to be parallel in content and which are reliable as defined by parallel forms criteria, KR-20, or kappa?

CHAPTER II

Basis of the Problem

Reliability is accepted to mean consistency by measurement experts; however, the means of estimating consistency are many. Since 1910, when Charles Spearman first used the term "reliability coefficient," which he defined as "the coefficient between one half and the other half of several measures of the same thing" (p.281), testing and measurement theoreticians have been working to develop estimates of reliability. These statistical estimates of reliability have been efficient attempts to show the measure of reliability that one would obtain if one could replicate administrations of an instrument to subjects time and time again. The purpose of this chapter is to present the concept of reliability; to present three statistical approaches that have been developed for estimating reliability; and to present the need, as expressed in the relevant literature, for a logical or empirical basis for estimating the reliability of criterion-referenced tests.

Concept of Reliability

Anyone interested in measuring something should be interested in the reliability of the measuring instrument. The instrument must produce the same or, at least, similar results when it is used to measure a thing again and again. If an instrument is extremely unreliable, decisions based on the results of its administration

might as well have been made without any measurement at all. To define this concept, reliability is considered to be the consistency of results which one obtains from replicating a measure time after time; this is often referred to as test-retest reliability. However, this procedure when applied in an educational environment may not be feasible, e.g., administration costs and student time spent in retesting rather than in other educational endeavors may be too great to warrant using this method. In addition, there are factors that confound the results obtained from this procedure, particularly when human subjects are being measured, i.e., individuals may remember in the second administration their responses in the first administration, thereby inflating the reliability coefficient; similarly, individuals' awareness of the area being tested might be heightened by taking the test the first time, so that, upon being retested, they would obtain higher scores, thereby deflating the coefficient. In other words, because resources are scarce, and because humans remember, grow, and change, even during the course of a week, the test-retest method of directly estimating reliability may often be impractical. Since the operational definition of reliability for this research cannot be directly measured, other statistical methods of estimating reliability will be considered to be indirect measures of what would be obtained if the test-retest method could be used.

Three Statistical Approaches to the Estimation of Reliability

The three statistical estimates of reliability which constitute the three levels of the independent variable for this study are ones that have been developed to estimate indirectly test-retest reliability.

In general, to compute a reliability coefficient, at least two test scores per subject in a group are needed. Two of the estimates of reliability in this study require two tests: parallel forms and κ. The other approach, KR-20, can be used with only one test, but it treats each item as if it were a parallel test to every other item. Both KR-20 and parallel forms, as reliability estimates, have long theoretical and empirical histories. Kappa, however, is a relatively recent development, put forth as a particularly appropriate estimate of the reliability of criterion referenced tests. The properties of each of these reliability estimates which are the three levels of the independent variable are discussed below.

"Parallel forms" is a term which is used freely in testing literature; however, to pinpoint a definition which is universally accepted in practice is extremely difficult. Criterion-referenced tests are often constructed to be parallel in content, in that many short criterion-referenced tests may be designed to measure one specific objective.

Basically, the accepted procedure for constructing such parallel tests is to 1) write items according to carefully stated specifications, and 2) select randomly the items for inclusion in the various tests. The resultant tests are then defined to be "randomly parallel" tests.

Another method for constructing parallel tests is to 1) construct two tests according to carefully stated specifications; 2) try them out on the same examinees; and 3) correlate the results. If the results are highly correlated, the tests are considered to be parallel.

Since the purpose of parallel forms is the estimation of the reliability that would be obtained if one were to use the test-retest method of directly measuring reliability, these methods are considered by some researchers to be expedient, rather than accurate estimates of reliability. Paul Horst (1966) recommended criteria for determining the existence of parallelism between forms of a test. The following were used for this study.

1. Each test has the same number of items.
2. The mean item difficulties are equal.
3. The dispersions of item difficulties are equal.
4. The variances of the item variances are equal.
5. The means of the inter-test item covariances are the same between tests as within tests.
6. The variances of the distributions of these covariances within and between tests are equal (pp.300-302).

Horst emphasized criterion of equality of the inter-test item covariances between tests. He maintained that without that equality of inter-test item covariances, theoretically, parallel tests could be composed of items from unrelated subject areas.

To summarize, although there are various criteria for parallel forms, the ones selected for this study were those developed by Paul Horst (1966) because they are the most clearly defined and specific. Parallel forms are one level of the independent variable.

The Kuder-Richardson Formula 20 is an estimate of reliability which assumes that every item has the same mean and the same variance; it treats every item as parallel to each of the other items. The Kuder-Richardson formula for the reliability of a test with n items will be

$$\frac{n}{n-1}\ \frac{s_t^2 - \Sigma p_i q_i}{s_t^2}\ , \qquad (1)$$

where t = test

s^2 = variance of the total test

r_{tt} = reliability of the test

n = number of items on the test

s_t^2 = variance of the total test

p_i = proportion of examinees who answered the i^{th} item correctly

q_i = proportion of examinees who answered the i^{th} item incorrectly

The Kuder-Richardson Formula 20 (KR-20) actually indicates the homogeneity or internal consistency within a test and is appropriate for use only when each item on a test is supposed to be measuring the same thing. It is an even more indirect method of estimating test-retest reliability than is the parallel forms method, in that it is, in a sense, an estimate of an estimate: it estimates the parallel forms reliability criteria. A simplified version of KR-20 is KR-21 which uses the mean item variance rather than the sum of all the item variances.

$$r_{tt} = \frac{n}{n-1} \cdot \frac{s_t^2 - n\overline{pq}}{s_t^2} \qquad (2)$$

KR-21 is often used in place of KR-20 because of the ease of its computation; however, in practice, KR-21 may actually be considerably smaller than KR-20 (Stanley, 1971) if item difficulties vary greatly (Ebel, 1972, pp.416, 418). KR-20 is an internal consistency coefficient between random parallel tests (Magnusson, 1966, p.117).

Formula KR-20 which is an internal consistency measure was selected to be the second level of the independent variable, because it is often used to estimate the reliability of criterion-referenced tests. An assumption on which KR-20 is based is that of equal item difficulties.

Kappa is a reliability estimate which measures the consistency of decisions across two forms of the same test or across repeated administrations of the same test. The coefficient κ is defined as

$$\kappa = (p_o - p_c)/(1 - p_c) , \tag{3}$$

where p_o, the observed proportion of agreement is given by

$$p_o = \sum_{i=1}^{k} p_{ii} , \tag{4}$$

where p_{ii} is the proportion of examinees placed in the i^{th} mastery state on both test administrations, and p_c, the expected proportion of agreement is given by

$$p_c = \sum_{i=1}^{k} p_{i.}p_{.i} . \tag{5}$$

In the last equation, $p_{i.}$ and $p_{.i}$ represent the proportions of examinees assigned to the mastery state i on the first and second tests, respectively. An example of the computation follows. The data are displayed in Table I.

Suppose that two tests have been administered to the examinees and they have been classified as masters and nonmasters as displayed in Table I. Then

$$p_o = .70 + .23 = .93$$

$$p_c = (.75)(.72) + (.25)(.28) = .61$$

$$\kappa = (.93 - .61)/(1 - .61) = .82$$

Table I

Example of the Contingency Table for Computation of Kappa

Test 2 / Test 1	Mastery States: Master	Mastery States: Non-Master	Marginal Proportions
Mastery States			
Master	.70 (.54)*	.05	.75
Non-Master	.02	.23 (.07)*	.25
Marginal Proportions	.72	.28	1.00

*Expected proportion

Kappa, which expresses the proportion of agreement on classification of masters and nonmasters on two tests beyond that expected by chance was the third level of the independent variable.

Estimating the Reliability of Criterion-Referenced Tests

In the context of the developing wide-spread use of criterion-referenced tests, a controversy has arisen in the tests and measurement area. This controversy centers on the issue of estimating the reliablity of criterion-referenced tests. As mentioned above, both the KR-20 and parallel forms estimates of reliability have long theoretical and empirical histories, but their histories

have been with norm-referenced tests. Their use or attempted use with criterion-referenced tests has led to a reexamination of the assumptions which underly their use in estimating reliability. The purpose of this section is to present the three theoretical positions which have been taken by measurement experts regarding criterion-referenced tests: 1) the traditional methods of estimating reliability are appropriate; 2) the traditional methods are appropriate but must be reinterpreted; and 3) the traditional methods are inappropriate and should be replaced. This last posture is the one which produced the suggestion that kappa be considered to replace the older methods. To establish the context for these three positions, a brief discussion of norm-referenced tests and criterion-referenced tests follows.

Norm-referenced and criterion-referenced tests differ primarily in that they are constructed to achieve different purposes. Norm-referenced tests are constructed to maximize variability of test scores. They are designed to make decisions in cases where one is interested in "'fixed quota' selection or ranking of individuals on some ability continuum" (Hambleton and Novick, 1973, p.162). Hambleton and Novick also point out that it would be possible to make criterion-referenced judgements and norm-referenced judgements about the results of either a norm- or criterion-referenced test, but because of the difference in purposes, and therefore, the differences in test construction procedures, neither norm-referenced judgements based on a criterion-referenced test,

nor criterion-referenced judgements based on a norm-referenced test would be particularly satisfactory (p.162).

The important issue for the purpose of this research is the fact that norm-referenced tests are constructed to maximize true variance among individuals, and that this variance, in relation to error variance, has been the basis in classical test theory, for developing the reliability estimates that have been used traditionally with norm-referenced tests. Because the concept of the relationship between true variance and error variance is inherent in the classical test theory concept of reliability, a brief discussion of the relationship follows. From the theoretical notations of classical test theory, the dependence upon variance of scores for estimating reliability can readily be seen. An individual's score on the f^{th} form of a test is defined as

$$X_{pf} = T_p + e_{pf} \tag{7}$$

where X_{pf} is the obtained score of the p^{th} person on the f^{th} form, T_p is the true score of that person, and e_{pf} is his/her error of measurement on that form. The variance of observed scores of a group can be represented by σ_x^2, the variance of true scores by σ_t^2, and the variance of errors of measurement by σ_e^2. If the magnitude of the error of measurement covaries zero with the magnitude of the true score, then

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2 \tag{8}$$

which indicates that the variance of the observed scores is equal

to the variance of the true scores plus the variance of the errors of measurement. Reliability of two parallel forms (or a test and a retest) has then been defined as

$$\rho_{ff'} = 1 - \frac{\sigma_e^2}{\sigma_x^2} \qquad (9)$$

A test is then defined to be unreliable in proportion to the magnitude of its error variance relative to its observed-score variance.

In summary, norm-referenced tests have been discussed with regard to their purpose, that of producing variance to produce information for "fixed quota" types of decisions. The reliability estimates for norm-referenced tests were shown to depend upon the relationship of true variance and error variance for their computation.

As stated above, criterion-referenced tests are constructed for different purposes than are norm-referenced tests. The operational definition of a criterion-referenced test for this research is the one developed by Glaser and Nitko (1971):

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

Measurements are taken in order to be able to make decisions about a student, or a group of students' mastery or nonmastery of a specified performance objective. Further, the performance objectives of interest to this research were developed by curriculum

specialists to be "minimal." Therefore, there was no quota on the number of students who could exceed the criterion, which was four out of five correct responses on the test. In fact, the educational goal is that of all of the students exceeding the criterion. It is at this point, all students exceeding the criterion, that the theoretical issue arises: there is no variability of scores. If reliability estimates must rely on variability of scores, then no estimate is possible. This is, of course, not the case, but this is the basis for the controversy.

To summarize, criterion-referenced tests have been discussed with regard to their purpose, that of measuring mastery or non-mastery of specific performance objectives to produce information for instructional purposes. Estimating the reliability of criterion-referenced tests, especially if all students "pass" them, may create a problem, because the traditional methods of estimating reliability have depended upon variance of test scores.

With this brief discussion of norm- and criterion-referenced tests, the basis for the controversy and the three theoretical positions which researchers have taken is established. Those researchers who believe that traditional methods are appropriate (Ebel, 1975; Livingston, 1972) maintain that lack of variability among examinees is, in fact, not a problem with criterion-referenced tests. Ebel (1975) says,

> The presumed need to redo test theory to accomodate criterion-referenced testing is probably based on mistaken assumptions.

> Even if test scores with little or no variability should be attained under the mastery learning model, which is seldom if ever the case, we can still test the effectiveness of the test, and of the items, by administering the test before and after instruction. The job of any achievement test, whether criterion-referenced or norm-referenced, is to differentiate among levels of achievement. To determine how well the test can differentiate, it must be given to examinees who have different levels of achievement. A group having these differences can always be found or assembled. Applied to the scores of individuals in such a group, classical measures of test reliability and item discrimination will indicate how well the test can do its job (p.85).

His position, then, seems to be that, in the real world, this theoretical lack of variability, in fact, does not exist.

Livingston (1972) developed a reliability coefficient from classical test theory which is based upon deviations from the criterion score, rather than from the mean. His definition of variance is the distance between the observed mean score and the criterion score. The farther from the criterion score the mean score falls, the greater the criterion-referenced reliability of the test for that particular group of examinees. Since the purpose of a criterion-referenced test is the determination of mastery or nonmastery status of students, Livingston's insistence on a continuum of ability, albeit differently defined, in order to adequately estimate the reliability of criterion-referenced tests, appears to dodge the issue (Harris, 1972). Hambleton and Novick (1973) suggest that Livingston missed the point for the criterion-referenced tests that are the subject of this study.

They state that the problem is "one of deciding whether a student's true performance level is above or below some cutting score" (p.168). In other words, the question is not "how far does the student's score fall from the criterion score?" but "is the student's true score above or below the criterion score?"

The position of those researchers (Hambleton and Novick, 1973; Millman, 1974) who have most strongly urged a reinterpretation of classical reliability coefficients has been effectively stated by Hambleton and Novick (1973). They state:

> It is well known from the study of classical test theory...that when the variances of test scores is restricted, correlational estimates of reliability and validity will be low. Thus, it seems clear that the classical approaches to reliability and validity estimation will need to be interpreted more cautiously (or discarded) in the analysis of criterion-referenced tests (p.167).

The replacement of traditional, classical test theory reliability coefficients with entirely different reliability estimates has been recommended by Hambleton and Novick (1972); Millman (1974); and Swaminathan, Hambleton and Algina (1974). These people have suggested that a possible replacement for the traditional reliability estimate would be kappa which measures the consistency of decisions across two randomly parallel criterion-referenced tests. (Randomly parallel tests are two sets of items which have been randomly drawn from a pool of items measuring an objective.) These researchers maintain that reliability could be defined as the consistency of decisions made about masters and nonmasters

across two randomly parallel forms of the same test. The coefficient kappa will reveal not only the consistency of decisions, but the proportion of consistent decisions exceeding those which could have been expected to occur by chance. One problem that arises with the reporting of the coefficient kappa, similar to that incurred by reporting Livingston's coefficient, is that of interpretation. Kappa is relatively new to the literature, and it is not intuitively easy to understand what a specific value of kappa means. Kappa can only reach +1 when there is perfect agreement along the main diagonal about masters and nonmasters across two or more forms of a test; otherwise, it is strongly affected by the marginal totals.

With the presentation of the three positions, the issues of interest to this research are clear:

1. If variance exists among the criterion-referenced test scores of the population of examinees, then the reliability estimates which have been developed from classical test theory should produce satisfactory estimates of reliability. Because these coefficients are known quantities, they are more meaningful to those who read them.

2. If the variability of the examinees is restricted due to the nature of the objectives being measured, then the reliability estimates of criterion-referenced tests will be deflated and will demand reinterpretation. In other words, the lower reliability coefficients of criterion-referenced tests must be weighted by some method to produce coefficients equivalent to equally reliable norm-referenced tests.

3. The third position would suggest the replacement of reliability estimates which depend upon vari-

> ability of scores with a reliability measure which depends upon variability of decisions about masters and nonmasters.

Each of these positions has its appeal and theoretical rationale. This research will attempt to provide some answers to the question: if these methods, traditional and new, were used with data obtained from administering criterion-referenced tests, would different decisions about the reliability of those tests be made on the basis of the two reliability estimates, Kuder-Richardson Formula 20 and Kappa.

CHAPTER III

Procedures

Since research and position papers have raised the issue of the feasibility of using traditional reliability indicators in estimating the reliability of criterion-referenced tests, this study was designed to examine two reliability estimates as they performed on empirical data collected from two populations of one state's students. The purpose of this section is the presentation of the procedures that were used to gather and analyze the data for this study. Part I defines the population of objectives from which the objectives were drawn which were measured by the test instruments. The population of test forms is also described in Part I, along with the data collection procedures. Part II describes the characteristics of the three samples of test form pairs which were drawn and the sampling procedures which were utilized to do so. Part III describes the data analysis procedures.

Part I, The Population of Objectives, Population of Test Forms, and Data Collection Procedures

Population of objectives. Since the basis for the construction of criterion- or objective-referenced tests is the objective or criterion which the tests are constructed to measure, a brief description of the population of objectives is essential. An entire population of objectives for an academic subject area is

a phenomenon of extraordinary size, difficult to conceptualize and probably not feasible to attempt to develop empirically. For those reasons, the objectives which were the partial basis for this study are described by their developers as "minimal" objectives for this subject area. They are minimal in the sense that they are assumed to be attainable by all students by termination of ninth grade. The objectives were developed by educators from various fields of specialization, including instruction, curriculum, measurement and research. In all, there exist approximately 500 objectives for grades kindergarten through nine in this subject area's population of minimal objectives.

For the purpose of this study, the objectives of interest are those which are to be attained in grades kindergarten through three and those which are to be attained in grades four through six. The former group was measured at the beginning of fourth grade and the latter at the beginning of seventh grade. However, not all of the approximately 400 objectives for grades kindergarten through six were measured by the tests which were used for this study. Only 29 objectives were measured in the fourth grade instrument and only 39 were measured in the seventh grade instrument. The reasons for this limitation were as follows:

1. Each objective was measured by a five item, multiple-choice, group administered test;
2. Objectives for other subject areas were also measured;
3. Students were allowed to take as long as necessary to complete the tests;

4. Testing time for the entire test battery was to be about three hours (although not in one day);

5. Only those objectives which could be measured by a paper and pencil, multiple-choice group administered instrument were desired.

Given these limitations, the objectives to be measured were selected by the testing staff with the advice of instructional and curriculum specialists according to the following criteria:

1. The importance of the objectives to the acquisition of future academic skills or to survival in the world outside that of academia; and

2. The feasibility of measuring the objectives by paper and pencil, multiple-choice, group administered instruments.

To summarize, the population of objectives consisted of 29 fourth grade minimal objectives and 39 seventh grade minimal objectives which described specific performances which 1) should be attainable by all students at the specified grade levels, 2) are important to future academic success or life survival, 3) could be administered in an approximate time interval (three hours), and 4) could be measured by paper and pencil, multiple-choice, group administered instruments.

Population of test forms. The total population of test forms consisted of 204 five-item tests. In all, there were 68 objectives to be measured, and three different test forms were used to measure each objective (68 objectives X 3 tests = 204 tests). One of the three tests measuring each objective was a

"core" test; the other two tests measuring each objective were "experimental" tests.

The 68 "core" tests, each measuring one objective, were tests which had been administered to students for three years (1973, 1974, 1975). In those years, a student was classified as master of an objective if four or five items measuring the objective were answered correctly; if a student answered fewer than four of the items correctly, the student was classified as a nonmaster of the objective.

The other two forms of each test were the "experimental" versions which were designed to measure the same objectives as were measured by the core tests. From a content standpoint, they were, in fact, each designed to be parallel measures to a core test of an objective. Subsequent tryouts and editing served to reinforce this design. Attainment and nonattainment for the experimental tests were defined in the same manner as had been done for the core tests.

In summary, the population of test forms consisted of three, five-item tests per each of the 68 objectives. One test for each objective had been administered three times to the population of students for which it was intended; the other two test forms were designated "experimental" and were designed to be parallel in content to the core tests. Attainment of an objective, as measured by these tests, was defined as correctly answering

four or five of the items. Nonattainment was defined as answering only three or fewer items correctly.

Data collection procedures. The 29 fourth grade core tests of the objectives were contained in a fourth grade core test booklet which was administered to the entire fourth grade population. The 39 seventh grade core tests of the objectives were contained in a seventh grade core test booklet which was administered to the seventh grade population. The experimental tests, however, were contained in separate booklets. Each booklet contained four tests which measured four different objectives. In order to obtain the sample of students who took each experimental booklet, the booklets were distributed by the method described below.

A spiral sampling plan was used to select the sample of students which took each experimental booklet. All students were administered the core booklet. Because the students recorded their answers for both the core and the experimental tests on the same answer sheet, each individual student's performance on the core test of an objective and on an experimental test of the same objective could be compared.

The spiral sampling was accomplished in the following manner. At the central distributing point, all of the experimental booklets were numbered from 1 to k; k indicates the number of the last booklet. Each experimental booklet contained 20 items measuring four objectives. A random number from 1 to k was selected, and

that random number became the first booklet, with the rest of the booklets following it in their sequential order. When k was reached, booklet number 1 followed it, and so on, until all of the booklets were ordered.

Because other experimental tests, besides those of interest to this study, were included, every student in the state who was a member of the fourth or seventh grade populations was in one of the samples taking an experimental booklet. This procedure provided a sample size of about 4000 students per booklet, and therefore, per objective and test form.

After the booklets had been arranged in the order described above, they were then counted out to be sent to the school districts, according to the number of students to be tested in each district. This number to be tested in fall, 1975, was obtained from the district in spring, 1975, and was the superintendent's projected fourth and seventh grade enrollment figures.

The district test coordinator was directed to count and distribute, without changing the experimental booklet order, enough booklets for each school. The school test coordinator was then directed to follow the same procedure for distribution to each classroom.

As specified above, every fourth and seventh grade student participated in the core tests. The only exceptions were certain special education students.

In summary, the data were collected by means of spiral sampling which is a form of systematic sampling. The sampling plan obtained a sample size of about 4000 students per test form pair. The data obtained from each sample were their responses to four core tests of four objectives and four matching experimental tests.

Units of analysis. The data collection procedures above led to three test forms from which two test pairs could be drawn, as shown in Figure 1. In each case, the experimental test of five items was paired in the student sample with the core test of five items. The number of test pairs per objective and the total number of test pairs are shown in Table II. In all, this procedure produced 136 test pairs, the units of analysis for the present research.

Table II

Test Pairs Obtained Per Objective

Number of Objectives		Test Pairs per Objective		Test Pairs
Fourth grade:	29	2		58
Seventh grade:	39	2		78
TOTAL:	68	X 2	=	136

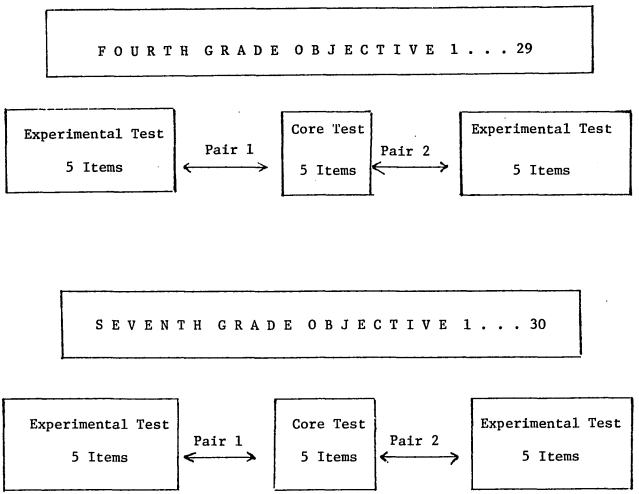

Figure 1. Method of obtaining two test pairs per objective.

Part II, Characteristics of the Samples and the Sampling Procedures

Characteristics of the samples of test forms. Since the purpose of this study was to examine different estimates of reliability, parallel forms, KR-20, and kappa, to find out whether or not each of the reliability measures produces different information about the reliability of these criterion-referenced tests, three samples of test forms were drawn from the population of pairs of test forms. Each sample consisted of 20 pairs of test forms. One sample represented those forms which, according to the traditional definition of "parallel," demonstrated the range of parallelism in the population of test pairs. The second sample represented the range of mean KR-20's in the population of test pairs. The third sample selected represented the range of kappas derived from this population of test pairs. These characteristics of parallelism, KR-20, and kappa constituted, operationally, the three levels of the independent variable.

All of the 136 pairs of tests were ranked from "very high" to "acceptable" on each of the three criteria for selection, parallel forms, mean KR-20's and kappa. They were then stratified into four strata of reliability: very high, high, moderate, and acceptable. From each stratum five pairs of tests were selected to be the sample on which the analysis was performed. The assumption behind this process was that the test development procedures were such that all of the tests would have at least "acceptable"

reliability. ("Acceptable," as specifically applied to each of the levels of the independent variable will be discussed below, since no general definition was appropriate.) If, in some cases, this assumption of acceptable reliability proved to be unfounded, unacceptable pairs were discarded. In Table III are shown the three levels of the independent variable and the strata from which the samples of test pairs were selected.

Table III

Sampling Plan for Stratified Random Selection of Test Pairs That Represent Characteristics of Parallel Forms, Mean KR-20 and Kappa

Reliability Coefficients	Level 1 Parallel Forms Test Forms	Level 2 Mean KR-20 Test Forms	Level 3 Kappa Test Forms
Very High	5	5	5
High	5	5	5
Moderate	5	5	5
Acceptable	5	5	5
TOTAL	20	20	20

Parallel forms sample. The first sample to be drawn was that which represented the range of parallel forms. As stated before, each triplet of test forms that measured an objective was written to be parallel in content. However, the traditional,

norm-referenced meaning of "parallel" is much more specific than human judgement about content; therefore, the following requirements for parallel forms were the basis for selection:

1. Each test had the same number of items (five).
2. The mean item difficulties were equal.
3. The dispersions of item difficulties were equal.
4. The variances of the item variances were equal.
5. The means of the inter-test item covariances were the same between tests as within tests.
6. The variances of the distribution of the covariances were the same between tests as within tests.

Since all of the tests consisted of five items, all test pairs were retained using the first criterion. The second step consisted of ranking all parallel forms on the basis of equal mean item difficulties (p-values). If p-values were within .02 points of each other (with rounding from the third decimal place), they were operationally defined to be "equal." These pairs were ranked on their mean p-values and were judged to be in the Very High stratum if the mean p-values differed between the two test forms by only .000 to .004; the High stratum if mean p-values differed from .005-.009; the Moderate stratum if mean p-values differed from .010-.014; and the Acceptable stratum if mean p-values differed from .015-.025.

The rationale for carrying the stringent definition of parallel forms only to the second criterion consisted of two main points. 1) the test pairs had been designed to be parallel in content and 2) an operational definition of "equal" for the remaining criteria would be an arbitrary one with little empirical evidence on which to base it. In this study, therefore, the obtained values for the other criteria were accepted to be "equal."

In all three samples, parallel forms, mean KR-20, and kappa sample, the number of cases which fell into a stratum was determined and each case was assigned an identification number, then a random number table was used to select the cases which would represent the stratum in the sample. For instance, if there were 18 members in the stratum, all numbers between 01 and 18 were eligible for selection. The cards on which the data were recorded were shuffled; then the random number table was used by the researcher so that the first five numbers between 01 and 18 which were encountered indicated the cards which were to be selected for the stratum of a given sample.

KR-20 sample. The second sample to be drawn was the sample representing the range of mean KR-20's. The mean KR-20 was obtained by adding together the KR-20 for the core test and the KR-20 for the paired experimental test and then dividing the sum by two (the number of test forms). The mean KR-20 was used in order to produce a coefficient which would be mutual to the test form pair, the unit of analysis. The Very High mean KR-20 stratum

was defined as those pairs having a mean KR-20 of .85 or higher; the High stratum consisted of those pairs having a mean KR-20 of .79-.84; the Moderate stratum consisted of those having a KR-20 of .73-.78; and the Acceptable stratum consisted of those with a mean KR-20 of .60-.72. In addition, if the KR-20's from the two tests differed by more than .10, they were not considered for inclusion in the sample, since they seemed to differ so much from one another that their values, taken separately, would have clearly put each one into a different stratum from the other.

The kappa sample. After the parallel forms sample and the KR-20 sample were selected, the kappa sample was randomly selected from the remaining test pairs, again according to the stratification plan outlined above. The remaining pairs were ranked from high to low according to the obtained values of kappa. Since kappa's obtained value is strongly affected by the marginal proportions, it is difficult to determine the range of specific values of kappa which may be obtained from the data. Kappa will reach 1 only if the decision making is perfect across both forms of the tests. Therefore, it was determined that a kappa of .30 define the lower limit on the Acceptable stratum. This would indicate, after the effects of chance agreement were removed, the proportion of joint agreements across the test. Intuitively, this seemed to be the lowest "acceptable reliability" that could be tolerated. Further, it was decided that the other three strata be defined as .40-.49, .50-.59, and .60-1.00. Because of a lack

of empirical evidence on the behavior of kappa, this plan seemed to be reasonable.

The dependent variables. Using the test pairs, the mean KR-20's and kappas were calculated. These two reliability coefficients were the dependent variables. Kappa, the coefficient which measures the consistency of decision making across two measures, produced one coefficient for each test pair. Obtaining the mean KR-20 for each test pair, however, involved two steps. 1) The KR-20 for each five-item test was calculated. This measured the homogeneity of the five items within each test. 2) The mean KR-20 for each test pair was calculated by adding together the KR-20 computed for each of the tests in the test pair and then dividing by two (the number of tests). This was done in order to have one KR-20 measure for each test pair, the unit of analysis.

Part III, Data Analysis

The research design involved the selection of three samples of 20 reliable test pairs. One sample was reliable according to parallel forms criteria; one according to mean KR-20 coefficients; and one according to the kappa coefficients. Parallel forms, KR-20, and kappa constituted the three levels of the independent variable, and the test pairs constituted the units of analysis. The mean KR-20 of each of the samples and the mean kappa of each of the samples were the two dependent variables. The null

hypotheses were that if the three indices provide the same information for decisions about the reliability of the tests, then the KR-20 coefficients for the three groups would be equal and the three kappa coefficients for the groups would be equal. The research hypotheses were that if either the KR-20's or the kappas for the three groups were different, then the reliability coefficients provide different information for decisions about the reliability of the test pairs.

Data analysis procedures. After the samples had been drawn, the dependent variables, the mean KR-20 and kappa, were analyzed using analysis of variance procedures. One ANOVA was performed to determine whether or not there was a difference among the mean values of KR-20 for the three groups of test pairs. The null hypothesis was that the mean KR-20 for the parallel forms level of the independent variable would be equal to the mean KR-20 of the KR-20 level of the independent variable which would be equal to the mean KR-20 of the kappa level of the independent variable.

$$H_o: \mu_{1,\ KR\text{-}20} = \mu_{2,\ KR\text{-}20} = \mu_{3,\ KR\text{-}20} \qquad (10)$$

where 1 = level of the independent variable representing the range of parallel forms,

2 = level of the independent variable representing the range of mean KR-20's,

3 = level of the independent variable representing the range of kappas.

Alternatively, the research hypothesis was that not all three mean KR-20's were equal.

$$H_1: \sum_{j=1}^{3} (\mu_{j,KR-20} - \mu_{\cdot,KR-20})^2 \neq 0 \qquad (11)$$

A second ANOVA was performed to determine whether or not there was a difference among the mean values of kappa for the three levels. The null hypothesis was that the mean kappas for the three levels would be equal:

$$H_o: \mu_{1,\kappa} = \mu_{2,\kappa} = \mu_{3,\kappa} \qquad (12)$$

The research hypothesis was that not all three mean kappas were equal.

$$H_1: \sum_{j=1}^{3} (\mu_{j,\kappa} - \mu_{\cdot,\kappa})^2 \neq 0 \qquad (13)$$

The Tukey method of comparing the difference between means was planned if the ANOVA procedures revealed differences at the .05 level of significance.

In summary, the data analysis procedures consisted of two ANOVAs: one performed using the mean KR-20's as the dependent variable for comparison among the three groups and one performed using the mean kappas as the dependent variable for the three groups. These procedures were consistent with the research problem which was to investigate whether or not these reliability estimates provided different estimates of the test-retest reliability of these criterion-referenced tests.

CHAPTER IV

Results

The purpose of this research was to provide empirical evidence, using criterion-referenced test results, of the behavior of two different reliability estimates; basically, to see whether or not each reliablity coefficient would provide a different estimate of reliablity. This chapter presents the results that were obtained using the procedures described in Chapter III. Part I provides a brief review of the units of analysis, the levels of the independent variable, and the dependent variables. Part II describes the characteristics of the samples that were selected Part III presents the results of the two analyses of variance procedures and the Tukey Post Hoc Comparison method.

Part I, Review of Units of Analysis, Independent Variable and Dependent Variables

The units of analysis of this study were test pairs. Each pair consisted of two sets of five items which had been constructed to measure a minimal performance objective. In all, there were 136 pairs. From the 136 pairs, three stratified samples were selected: the first sample represented the range of parallelism in the 136 pairs; the second represented the range of mean KR-20's; and the third represented the range of kappas. However, pairs which did not meet the criterion of "acceptable"

reliability as defined for each sample were excluded from consideration for that sample. Each sample included five randomly selected pairs in each of four strata. The four strata were defined, individually, for each sample and were acceptable, moderate, high and very high reliability estimates. These three samples constituted the three levels of the independent variable. The dependent variables were the mean KR-20 coefficients and the kappa coefficients for each test pair.

Part II, Characteristics of the Obtained Samples

Using the criterion for the parallel forms sample that the average p-values of the tests could differ by no more than .02 (with rounding), 89 of the 136 test pairs were eliminated, leaving 47 test pairs to be sampled. All other criteria of equality were assumed to be met. The pairs were ranked on their mean p-values and were judged to be in the Very High stratum if the mean p-value differed between the two test forms by .000 to .004; the High stratum mean p-values differed from .005-.009; and Moderate stratum mean p-values differed from .010-.014; and the Acceptable stratum mean p-values differed from .015-.025. This procedure produced 11 pairs in the Very High stratum, 12 in the High, 10 in the Moderate, and 14 in the Acceptable stratum. Shown in Table III are the ranges of differences which were obtained for each of the six criteria for parallel forms developed by Paul Horst (1966, pp.300-302).

Table IV

Range of Differences Between the Parallel Forms in the Four Strata

Strata	Criteria	Difference	
		Low	High
Very High Parallel (11)[a]			
	Mean p values	.001	- .003
	Variance of p values	.0	- .004
	Variance of item variances	.0	- .002
	Mean of covariance within	.005	- .019
	[illegible]e between	.011	- .042
	[illegible]es within	.0	- .0001
	[illegible] between	.0	- .0
High Parallel (1[illegible]		.005	- .009
		.0	- .014
	[illegible]ces	.0	- .003
	[illegible]thin	.0	- .023
	[illegible] between	.014	- .025
	[illegible]ovariances within	.0	- .001
	Variance of covariances between	.0	- .0004
Moderate Parallel (10)[a]	Mean p values	.010	- .014
	Variance of p values	.0	- .004
	Variance of item variances	.0	- .001
	Mean of covariance within	.011	- .023
	Mean of covariance between	.0122	- .045
	Variance of covariances within	.0	- .0007
	Variance of covariances between	.0	- .0006

Table IV
(con't)

Strata	Criteria	Difference	
		Low	High
Acceptable (14)[a]	Mean p values	.015	- .025
	Variance of p values	.001	- .003
	Variance of item variances	.0	- .002
	Mean of covariance within	.002	- .010
	Mean of covariance between	.014	- .038
	Variance of covariances within	.0	- .00
	Variance of covariances between	.0	- .0

[a]Numbers in parentheses indicate the available number of pairs in tests from which the five pairs that represented that stratum were selected.

Table IV

Range of Differences Between the Parallel Forms in the Four Strata

Strata	Criteria	Difference	
		Low	High
Very High Parallel (11)[a]			
	Mean p values	.001	.003
	Variance of p values	.0	.004
	Variance of item variances	.0	.002
	Mean of covariance within	.005	.019
	Mean of covariance between	.011	.042
	Variance of covariances within	.0	.0001
	Variance of covariances between	.0	.0
High Parallel (12)[a]	Mean p values	.005	.009
	Variance of p values	.0	.014
	Variance of item variances	.0	.003
	Mean of covariance within	.0	.023
	Mean of covariance between	.014	.025
	Variance of covariances within	.0	.001
	Variance of covariances between	.0	.0004
Moderate Parallel (10)[a]	Mean p values	.010	.014
	Variance of p values	.0	.004
	Variance of item variances	.0	.001
	Mean of covariance within	.011	.023
	Mean of covariance between	.0122	.045
	Variance of covariances within	.0	.0007
	Variance of covariances between	.0	.0006

Table IV
(con't)

Strata	Criteria	Difference	
		Low	High
Acceptable (14)[a]			
	Mean p values	.015 -	.025
	Variance of p values	.001 -	.003
	Variance of item variances	.0 -	.002
	Mean of covariance within	.002 -	.010
	Mean of covariance between	.014 -	.038
	Variance of covariances within	.0 -	.00
	Variance of covariances between	.0 -	.0

[a]Numbers in parentheses indicate the available number of parallel tests from which the five pairs that represented that stratum were selected.

The second sample to be drawn was that which represented the range of mean KR-20's, the second level of the independent variable. The Very High mean KR-20 stratum was defined as those pairs having a mean KR-20 of .85 or higher; the High stratum consisted of those pairs having a mean KR-20 of .79-.84; the Moderate stratum consisted of those having a KR-20 of .73-.78; and the Acceptable stratum consisted of those with a mean KR-20 between .60 and .72. In addition, if the KR-20's from the two tests differed by more than .10, they were not included in the sample, since they seemed to differ so much from one another that their values, taken separately, would have clearly put each into a different stratum. The Very High stratum contained 25 pairs, the High contained 14, the Moderate had 18, and the Acceptable contained 23 pairs, for a total of 80 pairs which met the KR-20 criteria and which had not already been selected for the parallel forms sample. Shown in Table V is the range of mean KR-20's obtained for each stratum, along with the criterion for selection that was used for each stratum.

The third level of the independent variable was the sample that represented the range of kappas. The Very High stratum of the kappa sample consisted of those pairs having a value of kappa between .60 and 1.00; the High stratum consisted of those with kappa of .50-.59; the Moderate, those pairs with a kappa of .40-.49; and the Acceptable stratum those pairs with a kappa of .30-.39. Nine test pairs were eligible for inclusion in the Very High stratum, 18 for the High, 24 for the Moderate, and 26 for the

Acceptable stratum. Shown in Table VI is the range of kappas for each stratum, along with the criterion for the section that was used for each stratum.

Table V

Range of KR-20's for the Four Strata

Group	Difference
	Low High
Very High Mean KR-20 (25)[a]	
Criterion for selection	.85 - 1
Obtained range	.85 - .910
High Mean KR-20 (14)[a]	
Criterion for selection	.79 - .849
Obtained range	.795 - .840
Moderate Mean KR-20 (18)[a]	
Criterion for selection	.73 - .789
Obtained range	.73 - .75
Acceptable Mean KR-20 (23)[a]	
Criterion for selection	.6 - .729
Obtained range	.675 - .705

[a]Numbers in parentheses indicate the available number of mean KR-20 pairs from which the five pairs that represented that stratum were selected.

Table VI

Range of Kappas for
the Four Strata

Group	Difference
	Low High
Very High Mean Kappa (9)[a]	
Criteria for selection	.6 -1.00
Obtained range	.62 - .73
High Mean Kappa (18)[a]	
Criteria for selection	.50 - .59
Obtained range	.52 - .56
Moderate Mean Kappa (24)[a]	
Criteria for selection	.4 - .49
Obtained range	.40 - .49
Acceptable Mean Kappa (26)[a]	
Criteria for selection	.3 - .39
Obtained range	.31 - .37

[a]Numbers in parentheses indicate the available number of kappas from which the five pairs that represented that stratum were selected.

To summarize, the characteristics of the obtained samples have been presented. The three samples met the criteria for selection; one represented the range of parallel forms; one represented the range of mean KR-20's; and one represented the range of kappas. The criteria for inclusion in the four strata within each sample were also met.

Part III, Results of the Analyses

Since the focus of this research was on the possible differences in estimation of reliability that the three methods might produce, an analysis of variance was performed on the KR-20 coefficients and the kappa coefficients, the dependent variables, for each of the three groups. The null and research hypotheses to be tested were:

$$H_o: \quad \mu_{1,KR\text{-}20} = \mu_{2,KR\text{-}20} = \mu_{3,KR\text{-}20} \tag{14}$$

$$H_1: \quad \sum_{j=1}^{3} (\mu_{j,KR\text{-}20} - \mu_{\cdot,KR\text{-}20})^2 \neq 0 \tag{15}$$

where 1 = level of the independent variable representing the range of parallel forms

2 = level of the independent variable representing the range of mean KR-20

3 = level of the independent variable representing the range of kappa

$$H_o: \quad \mu_{1,\kappa} = \mu_{2,\kappa} = \mu_{3,\kappa} \tag{16}$$

$$H_1: \quad \sum_{j=1}^{3} (\mu_{j,\kappa} - \mu_{\cdot,\kappa})^2 \neq 0 \tag{17}$$

where 1 = level of the independent variable representing the range of parallel forms

2 = level of the independent variable representing the range of mean KR-20

3 = level of the independent variable representing the range of kappa

In Table VII are shown the results of the ANOVA which addressed the question of whether or not there was a difference of the mean KR-20's for the three groups. As can be readily observed, the differences between the mean KR-20's for the three groups was not found to be significant at the .05 level. The three group means and standard deviations are shown in Table VIII.

In Table IX are shown the results of the ANOVA which addressed the question of whether or not there was a difference of the mean kappas for the three groups. As shown in the Table, there was a difference which was significant beyond the .05 level. The three group means and standard deviations are shown in Table X.

Table VII

Results of the ANOVA to Test the Hypothesis that the Means of the Mean KR-20's for Three Groups Were Unequal

Source	Sum of Squares	df	Mean Square	F	Probability
Between	.02	2	.01037	1.311	.28
Within	.45	57	.0079		
Total	.47	59			

Table VIII

The Means and Standard Deviations
of the Mean KR-20 for Each
Level of the Independent Variable

	Size	Mean	Standard Deviation
1, Parallel Forms	20	.740	.0888
2, KR-20	20	.786	.0780
3, Kappa	20	.761	.0987

Table IX

Results of the ANOVA to Test the
Hypothesis that the Mean Kappas
for Three Groups Were Unequal

Source	Sum of Squares	df	Mean Square	F	Probability
Between	.158	2	.07899	5.728	.0054
Within	.786	57	.01379		
Total	.944	59			

Table X

The Means and Standard Deviations of the Kappas for Each Level of the Independent Variable

Group	Size	Mean	Standard Deviation
1, Parallel Forms	20	.3815	.1174
2, KR-20	20	.4750	.1036
3, Kappa	20	.5010	.1297

In order to detect which of the group means was producing the significant results, the Tukey method for comparing the differences between means was used. The results are shown in Table XI.

Table XI

Results of the Tukey Post Hoc Method of Comparing the Differences Between the Kappa Means for the Three Groups

Population Contrast	Sample Constrast	Confidence Interval*	Significant
$\mu_{1,\kappa} - \mu_{2,\kappa}$	-.0935	-.0039, - .1831	yes
$\mu_{1,\kappa} - \mu_{3,\kappa}$	-.1195	-.0300, - .2091	yes
$\mu_{2,\kappa} - \mu_{3,\kappa}$	-.0260	-.1156, + .0636	no

*$1 - \alpha\, q_{j,j(n-1)}(MSw/n) = (3,405)(.0263) = .0896$

The Tukey method thus revealed that both group 2, the mean KR-20 group, and group 3, the kappa group, differed from group 1, the parallel forms group. However, groups 2 and 3 did not differ from each other.

To summarize, the results of the ANOVA to test the difference between the means of the mean KR-20's for the three groups revealed no significant differences at the .05 level. The ANOVA to test the difference between the mean kappas for the three groups showed a significant difference beyond the .05 level. The Tukey method for post hoc comparisons showed that both the KR-20 group mean and the kappa group mean differed from the parallel forms group mean. The null hypothesis of equal means for the KR-20 group and the kappa group was not rejected.

CHAPTER V

Discussion and Conclusions

The purpose of this research was to examine three reliability estimates to find out whether or not they would produce different estimates of the reliability of a set of criterion-referenced tests. Three samples were selected, representing the range of acceptably reliable parallel forms, mean KR-20 reliability coefficients, and kappa coefficients. Analysis of variance procedures were used to test whether or not the mean KR-20's for the three groups were different and to test whether or not the mean kappas for each group were different. The finding was: 1) the mean of the kappas did differ beyond the .05 level of significance.

The purpose of this chapter is to discuss those results, to present some tentative conclusions, and to suggest further research, Part I discusses the results of the two ANOVA's in relation to the resultant samples which were analyzed. Part II presents the conclusions. Part III describes further research which needs to be done.

Part I, Sample Selection Procedures and the ANOVA's

Parallel forms sample. As indicated in the procedures section, the parallel forms sample was selected from the test

pairs that represented an acceptable range of parallelism. The criteria for parallelism were those developed by Horst (1966):

1. Each test had the same number of items.
2. The mean item difficulties were equal.
3. The dispersions of item difficulties were equal.
4. The variances of the item variances were equal.
5. The means of the inter-test item covariances were the same between tests as within tests.
6. The variance of the distributions of these covariances were the same between tests as within tests.

In fact, the tests which were selected to represent the range of acceptable parallelism were selected on the basis of criteria (1) and (2). Two facts accounted for this; 1) "Equal" for criteria (2) through (5) was difficult to define. For the case of criterion (2), the mean item difficulties were equal, a difference of no more than .02 (with rounding) between two p-values was tolerated. 2) Only 47 of the 136 test pairs met this criterion. For criteria (3) through (5), all differences were tolerated, and the criteria were considered to be met. Table IV presented the differences that were obtained for the various criteria, and it can be seen that they ranged from 0 to .045. Further, the largest differences that were obtained were those described by criterion (5), the means of the inter-test item covariances were the same between tests as within tests. The differences that were obtained are shown in Table XII.

Table XII

Range of Differences for Mean of Covariances Between and Within Tests for the Four Strata of the Parallel Forms Sample

	Very High	High	Moderate	Acceptable
Covariance Within	.005 - .019	.0 - .023	.011 - .023	.002 - .010
Covariance Between	.011 - .042	.014 - .025	.012 - .045	.014 - .038

The differences in the covariances between the tests are pointed out for special consideration, because they represented the largest differences that were found, and they can be considered to go beyond the criteria necessary to estimate reliability. Horst points out one could construct tests measuring numerical ability, spatial ability and verbal ability and meet criteria (1) through (4). In other words, one could reliably estimate test-retest reliability if parallel forms met those four criteria. Horst (1966), however, goes one step further in demanding that the mean covariances between and within test be equal. He states:

> Obviously, if we impose this further condition we do not get the unacceptable state of affairs where an arithmetical, a verbal, and a spatial test can be parallel forms of the same test... Clearly, the average covariance of arithmetic

> items would not be the same as the average covariance between arithmetic and space relations or verbal items. These latter average covariances would in general be less (p.302).

In other words, the criterion that the mean covariances between and within tests should be equal is one that relates to validity rather than to reliability. If one is willing to accept this conclusion, then the parallel forms sample did meet the criteria of parallelism reasonably well. No differences greater than .023 existed in the sample.

Mean KR-20 sample. After the parallel forms sample was selected, 80 test pairs representing the range of acceptable mean KR-20's were available for sampling. As indicated in the Procedures section, mean KR-20's were obtained in order to have one coefficient for each test pair. For the KR-20 sample, the KR-20's for the tests comprising a pair were not allowed to differ by more than .10, since a difference of that magnitude would have clearly put the separate KR-20's into two different strata. This last criterion was used for the selection of the KR-20 sample, however, but did not apply to the mean KR-20's which were computed for the parallel forms sample nor for the kappa sample. For this reason, which will be discussed more thoroughly below, information was probably lost about the KR-20's of the other two samples. For example, if one of the test pairs in either of the other samples had one KR-20

of .90 and one of .40, the resultant mean KR-20 would be .65 which would not be very reflective of either of the real values. In fact, differences in the two KR-20's greater than .10 occurred in four pairs in the kappa sample and in six pairs in the parallel forms sample. To correct this problem, the KR-20 could have been computed for all 10 items of the two tests which comprised each of the test pairs.

Kappa sample. After the parallel forms sample and the mean KR-20 sample were selected, there remained 77 test pairs representing the range of acceptable kappas. No problems existed in the definition or selection of this sample.

To summarize, the parallel forms sample did meet the criteria of equality across 1) average p-values; 2) variance of p-values; 3) variance of item variances; 4) mean of covariance within tests. Although differences as large as .045 existed between the means of the covariances between tests, this criterion was considered to be of minimal importance because of its relationship to validity rather than reliability. The sample of parallel forms was considered to be representative of parallel forms. The mean KR-20 sample was representative of the KR-20's. Some loss of information about the KR-20's of the other two samples probably occurred, because of the potential averaging of KR-20's of extremely different magnitudes. No problems were noted with the kappa sample.

Analysis of variance of the mean KR-20's for each group. As shown in Table VII the difference in the mean KR-20's for the three

groups was not found to be significant at the .05 level of significance. Two conclusions are possible. The effect of using the mean of the two KR-20's for each test pair removed the differences that may, in fact, exist. If there are differences in the KR-20's for the three groups, they would more likely show up if the KR-20 had been computed for the 10 item set, rather than separately for each five item set. The other conclusion that is possible is that tests which are reliable according to parallel forms, KR-20 or kappa criteria will produce the same KR-20 estimates of test-retest reliability. The position which states that there may be a need to reinterpret the reliability coefficients for use with criterion-referenced tests would be upheld in the sense that obtained values of KR-20 for criterion-referenced tests may be lower, but the criterion-referenced tests are probably just as reliable as a norm-referenced test reporting a higher value of KR-20. The mean KR-20 of the three groups of reliable tests ranged from .74 to .79 which is somewhat less than one would require for norm-referenced tests.

Analysis of variance of the kappas for each group. As shown in Table IX the difference of the mean kappas for the three groups was found to be significant beyond the .05 level. In order to find out where the differences existed, the Tukey Post Hoc method for comparing the differences among means was used. As indicated in Table X, both the KR-20 sample and the kappa sample differed from the parallel forms sample, but they did not differ from each

other. Since these results were more conclusive that those obtained from the ANOVA performed on the mean KR-20's, they will be discussed in Part II.

Part II, Conclusions

The careful development of parallel forms according to the criteria set forth by Paul Horst would seem to be effort wasted in the case of short, criterion-referenced tests. For example, if one had used the criteria for parallel forms to be the determination of the reliability of these test pairs, only 77 of the 136 test (59%) pairs could have been utilized. Further, a user of those parallel tests would have been disappointed in the results, since the consistency of decisions about examinees for those tests was less than that for either the parallel forms sample or the KR-20 sample. These criteria for parallel forms need further investigation with longer, norm-referenced tests to determine their usefulness and also to determine the range of values for each of the criteria which could be interpreted as being "equal."

Two further conclusions, based upon the practical, can be offered. If one has developed randomly parallel criterion-referenced test, i.e., a set of items measuring a specific objective which have been randomly assigned to two or more tests, then the decision to be made is whether to use KR-20 or kappa. KR-20 is the more commonly reported and its qualities are more

familiar to researchers. For those reasons, it may be preferred to kappa. However, if, as these data show, and theoreticians have maintained, obtained values of KR-20 need to be reinterpreted in the context of criterion-referenced testing, then KR-20 may hold little advantage over kappa. Kappa is readily interpretable as the consistency of decisions made across two tests, after the effects of chance have been removed. Particularly at the classroom level, kappa is more easily computed than is KR-20. The research audience needs to develop an intuitive feeling for the meaning of the various values which kappa can aquire. Further research needs to be done in order to establish what levels of kappa are desirable. This will be discussed in more detail in Part III.

In summary, the conclusions derived from this study are the following: 1) if one is interested in developing parallel criterion-referenced tests, using the parallel forms criteria described by Paul Horst (1966, pp.301-303) requires too great an expenditure of energy for the value received; 2) there is not enough evidence from this study to indicate that consistency of decisions about masters and nonmasters across parallel tests, as measured by kappa, would indicate homogeneity of items within or between tests; 3) both kappa and KR-20 have advantages and disadvantages in the reporting of their values.

Part III, Suggestions for Further Research

As indicated above, further research needs to be done using the KR-20 coefficient computed across the items of two or more randomly parallel tests and the kappa computed for those tests, in order to determine the relationship between KR-20 and kappa. The question to be addressed is, "Does one have certain knowledge of the consistency of decisions that will be made across two or more tests, if one has knowledge of the level of homogeneity (KR-20) of those two tests?"

The other large topic for future research is that of the behavior of kappa under varying conditions. What level of kappa is desirable, in order to say that two tests are reliable in the consistency of decisions that they produce?

REFERENCES

Bohrnstedt, G. W. Reliability and Validity Assessment in Attitude Measurement. In G. F. Summers (Ed.), Attitude Measurement. Chicago: Rand McNally, 1971.

Cohen, J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 1960, 20, 27-46.

Ebel, R. L. Educational Tests: Valid? Biased? Useful? Phi Delta Kappan, 1975, 57, 83-88.

Ebel, R. L. Essentials of Educational Measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1972.

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

Glass, G. V., & Stanley, J. C. Statistical Methods in Education and Psychology. Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

Hambleton, R. K., & Novick, M. R. Toward an Integration of Theory and Method for Criterion-Referenced Tests. Journal of Educational Measurement, 1973, 12, 159-170.

Harris, C. W. An Interpretation of Livingston's Reliability Coefficient for Criterion-Referenced Tests. Journal of Educational Measurement, 1972, 9, 27-29.

Horst, P. Psychological Measurement and Prediction. Belmont, California: Wadsworth, 1966.

Livingston, S. A. Criterion-Referenced Applications of Classical Test Theory. Journal of Educational Measurement, 1972, 9, 13-26.

Magnusson, D. [Test Theory] (H. Mabon, trans.). Reading Massachusetts: Addison-Wesley, 1966.

Millman, J. Criterion-Referenced Measurement. In W. J. Popham (Ed.), Evaluation in Education. Berkeley: McCutchan, 1974.

Spearman, C. Correlation Calculated from Faulty Data. British Journal of Psychology, 1910, 2, 271-295.

Stanley, J. C. Reliability. In R. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Swaminathan, H., Hambleton, R., & Algina, J. Reliability of Criterion-Referenced Tests: A Decision-Theoretic Formulation. Journal of Educational Measurement, 1974, 11, 263-266.