ED 091 432 TM 003 646

AUTHOR	Rim, Eui-Do; Bresler, Samuel
TITLE	Livingston's Reliability Coefficient and Harris' Index of Efficiency: An Empirical Study of the Two Reliability Coefficients For Criterion-Referenced Tests.
PUB DATE	[74]
NOTE	21p.; Paper presented at a Joint Session of the American Educational Research Association and the National Council on Measurement in Education (Chicago, Illinois, April 1974)
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	Academic Achievement; *Correlation; *Criterion Referenced Tests; Elementary School Mathematics; Post Testing; Pretesting; *Statistical Analysis; *Test Reliability; True Scores
IDENTIFIERS	Harris Index of Efficiency; Kuder Richardson Reliability Coefficient; Livingstons Reliability Coefficient; Variance (Statistical)

ABSTRACT

Livingston's reliability coefficients and Harris' indices of efficiency were computed along with the classical internal consistency coefficients, KR-20's (Kuder-Richardson internal consistency coefficient), for 678 criterion-referenced tests in the A through E levels of an individualized mathematics program. The coefficients were carefully studied and compared with each other in relation to the number of students, the number of items, the percentage points of the mastery criterion score and the mean, the absolute value of difference of the mean from the mastery criterion score expressed both as a percentage and in a standard score form, the standard deviation, the proportion of mastery students, the shape of the score distribution, and the mastery status indices derived from the cross-tabulated tables of students' performance on the pretest and the Curriculum Embedded Test (CET), the pretest and the posttest, and the CET and the posttest. (Author/RC)

17.05

Livingston's Reliability Coefficient and Harris' Index of Efficiency:
An Empirical Study of the Two Reliability Coefficients
for Criterion-Referenced Tests

Eui-Do Rim

Research for Better Schools, Inc.

Samuel Brezler

AERA & NCME

Chicago, Illinois

1974

ABSTRACT

Livingston's reliability coefficients and Harris' indices of efficiency were computed along with the classical internal consistency coefficients, KR-20's, for 678 criterion-referenced tests in the A through E levels of IPI Mathematics, Edition II. The coefficients were carefully studied and compared with each other in relation to the number of students, the number of items, the percentage points of the mastery criterion score and the mean, the absolute value of difference of the mean from the mastery criterion score expressed both as a percentage and in a standard score form, the standard deviation, the proportion of mastery students, the shape of the score distribution, and the mastery status indices derived from the cross-tabulated tables of students' performance on the pretest and the Curriculum Embedded Test (CET), the pretest and the posttest, and the CET and the posttest.

INTRODUCTION

Two procedures have recently been proposed for the estimation of the reliability of a criterion-referenced test from total test scores.

Livingston (1970) derived a reliability coefficient for a criterion-referenced test by redefining the variance as a deviation from the mastery criterion score rather than from the mean score as it is in the sense of classical test theory. He showed the relation between the classical reliability coefficient and his reliability coefficient for criterion-referenced test, $K^2(X,T)$, as:

$$K^2\ (X,T) = \frac{\rho^2\ (X,T)\ \sigma_X^{\ 2} + (\mu_X - C)^2}{\sigma_X^{\ 2} + (\mu_X - C)^2} \qquad (1)$$

where ρ^2 (X,T) is a classical reliability coefficient, $\sigma_X^{\ 2}$ is the test variance, μ_X is the test mean and C is the mastery criterion score.

Livingston's proposal has been subjected to a substantial amount of critical analysis: Hambleton and Novick, 1972; Shavelson, Block, and Ravitch, 1972; Harris, 1972-a; and Raju, 1973. The primary criticism within these analyses centered around the inclusion of the $(\mu\text{-c})^2$ term. Specifically, Shavelson, Block and Ravitch (1972) observed that the term $(\mu\text{-c})^2$ dominates in deciding $k^2(X,T)$ for the criterion-referenced test where the test variance is relatively small. Hambleton and Novick (1972) indicated that Livingston's coefficient misses the essential point of criterion-referenced testing, and that the critical problem is one of deciding whether a student's true score is above or below the mastery criterion score, not one of showing how far his obtained score departs from the criterion score. Harris (1972-a) and Raju (1973) independently derived the same formula through the utilization of the two groups approach, under different assumptions, and concluded that Livingston's coefficient was impractical and unreasonable because it seemed to hardly meet their assumptions. In addition, Harris (1972-a) also stated that "although Livingston's reliability coefficient is generally larger than the conventional one, the standard error of measurement (which gives more meaningful information in deciding whether the student has a true score below or above a certain mastery criterion score) is the same."

At the 1972 AERA Meeting in Chicago, Harris (1972-b) proposed his index of efficiency:

$$\mu c^2 = \frac{SS_b}{SS_b + SS_w} \quad , \qquad (2)$$

where SS_b and SS_w denote the between- and within-group sums of squares that are determined by the two groups resulting from the dichotomization into mastery and non-mastery categories. Technically, his index of efficiency represents the correlation between the dummy variable that designates the group (mastery or non-mastery) and the total test score. Therefore, it does not depend upon the number of items. In this sense, it differs from conventional coefficients which increase as the number of items increases. It is, however, similar to them in dropping to 0.00 when all or none of the tested students belong to the mastery group. In addition, the index becomes 1.00 when the following conditions are satisfied: (1) the students are divided into mastery and non-mastery groups, and (2) the within-group variance is equal to zero. As an extreme case, the index is 0.00 when all the students achieve above the mastery criterion score. It changes to 1.00 when even a student misses one item on a 5 item test which has 100% correct response as the mastery criterion score. Marshall (1973) made an intensive study on the behaviors of Harris' index with simulated data. Among his findings that relate to the present study are: (1) the index is not affected significantly by either the number of subjects or by the number of items, (2) the index is affected by changes in the criterion; the higher the criterion, the lower the index, except when the total scores are all close to the number of items, in which case the trend is reversed, (3) the index increases as the range of competence increases for a given category of input competence vector, (4) the index decreases when the unaccounted for error variance increases, except when total scores are for the most part well above the criterion level, and (5) the index is generally higher as the mean of the test increases, for a given criterion level, unless the total score distribution is high in the extreme.

The present study intends (1) to investigate the behaviors of the two coefficients and the conventional reliability coefficient (KR-20) computed on the basis of real data that were collected from three I.P.I. Mathematics Edition II field test schools in relation to the number of students (N), the number of items (K), the percentage point of the mastery criterion score (Pc) and the mean ($P\bar{x}$), the absolute difference of the mean from the mastery criterion score expressed in percent ($|P\bar{x} - Pc|$) and in a standard score form ($|\bar{x}-c|/SD$), the standard deviation (SD), the percent of mastery students (Pm), the test type (Pretest, Curriculum Embedded Test, and Posttest), and the shape of the score distribution (normal, J-shaped, L-shaped, rectangular, etc.); (2) to compare the average size of the two coefficients for each level of the factors mentioned in (1); and (3) to study the relation of the two coefficients to the mastery status indices derived from cross-tabulated tables of students' performances on pretest and CET, pretest and posttest, and CET and posttest.

It is hoped that the present study will yield useful, significant information which might aid the development of theory and the improvement of practice in criterion-referenced testing.

DATA, METHODS AND PROCEDURE

The data used in the present study were collected from three IPI Mathematics, Edition II field test schools in 1971-72 school year. The IPI Mathematics, Edition II is a new version of IPI Mathematics which was originally developed by Learning Research and Development Center of University of Pittsburgh, revised by Research for Better Schools, and published by Appleton-Century-Crofts. It covers K-6 contemporary mathematics content which is divided into 10 content areas; Numeration and Place Value, Addition and Subtraction, Multiplication, Division, Fractions, Money, Time, Systems of Measurement, Geometry, and Applications. Instructional objectives in each content area are grouped into several levels (mostly Level A through Level G).

The student who is placed in an appropriate level on the basis of his or her placement test score takes the pretest which consists of items designed to measure the terminal behavior(s) of each objective in the unit. The student begins his study with the lowest numbered skill in the unit on which he did not demonstrate mastery on the pretest. Right after the lesson, the student takes the Curriculum Embedded Test (CET). If the student shows mastery on the CET, he then moves to the next unmastered skill. When the student completes all of the unmastered skills in the unit, he then takes the unit posttest. Therefore, the CET's can be regarded as immediate posttests and the posttests as delayed ones. These tests were administered on an individual basis. Consequently the number of students who took the test varies from test to test.

A computer program named SCOREWT3 was specially developed for the purpose of this study. It provides the user with a score distribution, mean, median, standard deviation, coefficient alpha of which KR-20 is a special case, and Livingston's coefficient and the proportion of mastery students when a mastery criterion score, C, is specified. It also gives Harris' Index of Efficiency, μc, and μc^2 for each of available score points in the score distribution upon user's request.

Thus far, 274 A-E level pretests, 209 A-D level CET's, and 212 A-D level posttests have been analyzed. Nine pretests, one CET and seven posttests were not used as data because they were one-item tests. The actual number of tests that constitute the data of the present study is presented in Table 1.

Table 1. Number of Test Data

Level	Test Type	CONTENT AREA										TOTAL
		N/PV	A/S	Mult.	Div.	Fract.	Money	Time	SOM	Geom.	Appl.	
A	Pre	13*	16*	-	-	3	1	0*	-	-	-	33
	CET	13	17	-	-	3	1	0	-	-	-	34
	Post	13*	16*	-	-	3	0	0	-	-	-	32
B	Pre	6*	12	4	3	3	1	1*	3	3	3	39
	CET	7*	12	4	3	3	1	1	3	3	3	40
	Post	6*	12	4	3	3	0	1*	3	3	3	38
C	Pre	14	13	7	4	6	5	6	6	1	8	70
	CET	13	13	7	4	6	5	5	6	1	8	68
	Post	14	13	7	4	6	5	6	6	0	8	69
D	Pre	5	10	9	7	7	5	4	6	4	9	66
	CET	5	10	9	7	7	5	4	6	4	9	66
	Pre	5	10	9	7	7	5	4	6	4	9	66
E	Pre	6	4	7	9	11	-	4	5	6	5	57
TOTAL	Pre	44	55	27	23	30	12	15	20	14	25	265
	CET	38	52	20	14	19	12	10	15	8	20	208
	Post	38	51	20	14	19	10	11	15	7	20	205

* One, two or three one-item tests were excluded from the unit.

The test consistency index and/or the efficiency index of instruction were derived from the results of the cross-tabulation of two test scores as follows:

		FIRST TEST	
		Non-mastery	Mastery
SECOND TEST	Mastery	Pnm-m *	Pm-m
	Non-mastery	Pnm-nm	Pm-nm

* The P's in the table represent the percentage.

$$I_{\text{Pre-CET}} = \text{Pm-m} + \text{Pnm-m} - \text{Pm-nm} \qquad (3)$$

$$I_{\text{Pre-Post}} = \text{Pm-m} + \text{Pnm-m} - \text{Pm-nm} - \text{Pnm-nm} \qquad (4)$$

$$I_{\text{CET-Post}} = \text{Pm-m} - \text{Pm-nm} \qquad (5)$$

All reliability and other information for a test were recorded on a standard optical scanning sheet from which the data card was punched. Since it was impossible to make a negative sign on the standard optical scanning sheet, the negative values of KR-20's and Livingston's coefficients were recorded as 0's.

Correlations were computed by BMD03D (Dixon, 1970) for pretest, CET, and posttest data separately and then for the combined total test data.

Data were grouped into 2 - 4 categories according to the frequency listing of the number of cases (N), number of items (K), percentage points of mastery criterion score (Pc) and mean ($P\bar{x}$), the difference between the mean and the mastery criterion score expressed in both percentage ($|Px - Pc|$), and standard score form ($|\bar{x} - c|/SD$), standard deviation (SD), the proportion of mastery students (Pm), and the shape of score distribution (SSD). Then nine two-factor multivariate analyses of variance were performed in order to compare the magnitudes of KR-20's, Livingston's coefficients, and Harris' maximum μc^2's and μc's. The first three-level factor was the same for all MANOVA's: test type; pretest, CET and posttest. The second factor in each of the MANOVA's consisted of one of the above mentioned variables blocked into two to four categories. The dependent measures in each MANOVA were the four coefficients; KR-20, $K^2(X,T)$, maximum μc^2, and μc. MANOVA was used in order to perform 4 ANOVA's at the same time. Prior to MANOVA, KR-20's, Livingston's coefficients and Harris' indices were transformed into Fisher's Z's, and Harris' μc^2's were converted into radians by arcsine transformation following Edwards' (1968) recommendations.

Only the results of the correlational study are reported in this paper. The results of MANOVA will be presented in a separate paper.

RESULTS

The cross tabulation results revealed that the distributions of KR-20's, Livingston's coefficients, and Harris' indices were quite different for the pretests, CET's and posttest's (χ^2 = 156.38 with 20 d.f.s for KR-20, χ^2 = 127.47 with 20 d.f.s for K^2 (X,T), and χ^2 = 48.14 with 14 d.f.s for μc's). Generally pretest coefficients showed negatively skewed distribution with fewer extreme values (such as 0.0 and 1.00). The distributions of CET's and posttests were less skewed than that of pretests, but there were more extreme values, especially 0.0 values.

The correlation of test type (value 1 was assigned to pretests, 2 to CET's and 3 to posttests) with KR-20, K^2(X,T) and μc were -.27, -.26 and -.04, respectively, with the first two coefficients being significant at the .01 level. The -.04 value was not significant. The difference between the last two coefficients was statistically significant at the .01 level when Hotelling's t-test (Walker & Lev, 1953, 259-260) was applied (t = 5.25). The results imply that larger KR-20 and K^2(X,T) coefficients are obtainable when a CRT is used as a pretest. Meanwhile, μc does not change much along with the shift in test-type. The results seem quite reasonable if the fact that greater test variance may be expected when a test is used as a pretest than when used as a CET or as a posttest is taken into consideration, and also that the μc does not have any relation with the variance. Therefore, further analyses were carried out for pretests, CET's and posttests separately hereafter.

A. Means, Standard Deviations and Intercorrelations of the Three Reliability Coefficients.

Table 2 presents the means and standard deviations of KR-20, K^2(X,T) and μc for pretests, CET's, posttests and for the combined total test data. The significance of mean difference between K^2(X,T) and μc was tested by using the t-test technique for paired observations (Walker & Lev. 1953, 151-154).

TABLE 2. Means and Standard Deviations of the Three Reliability Coefficients

Test Type		KR-20	$K^2(X,T)$	μc	t
Pretest (N=265)	Mean SD	.730 .237	.822 .188	.835 .104	.92
CET (N=208)	Mean SD	.415 .299	.628 .270	.755 .295	3.92**
Posttest (N=205)	Mean SD	.542 .309	.673 .259	.818 .213	5.37**
TOTAL (N=678)	Mean SD	.577 .309	.717 .252	.805 .214	6.22**

** Significant at the .01 level.

TABLE 3. Intercorrelations

Pretest

	KR-20	$K^2(X,T)$
$K^2(X,T)$	.838	
μc	.124	-.164

CET

	KR-20	$K^2(X,T)$
$K^2(X,T)$	.505	
μc	.445	-.364

Posttest

	KR-20	$K^2(X,T)$
$K^2(X,T)$	.684	
μc	.318	-.339

Total

	KR-20	$K^2(X,T)$
$K^2(X,T)$	.702	
μc	.359	-.246

On the average, the μc mean was higher than the $K^2(X,T)$ mean for all the pretest, CET, and posttest cases. The mean difference was significant at the .01 level for the CET, posttest, and the combined data. The mean difference for pretest was not statistically significant, but the standard deviation of μc's was considerably smaller than that of $K^2(X,T)$. As was expected, the mean of $K^2(X,T)$ was always higher than that of KR-20 for all test types. KR-20 had the largest standard deviation among the three coefficients for all test types.

Intercorrelations between two of the three reliability coefficients are presented in Table 3. All correlation coefficients are statistically significant at the .01 level except for the correlation between KR-20 and μc based on the pretest data which is significant at the .05 level. The KR-20 and $K^2(X,T)$ coefficients derived from the pretests were very highly correlated which seems to imply that the pretest situation is quite similar to a classical testing situation, insofar as these coefficients are concerned. It is worthwhile to notice that the two reliability coefficients for a criterion-referenced test are negatively correlated across all of the test types.

B. Influence of Related Variables on the Three Reliability Coefficients

It is very difficult to single out the effects of any one variable on the three reliability coefficients, because they all have more than two terms in their respective computational formulae and each variable is interdependent with many other variables and conditions. In this section, the zero-order correlations of the three coefficients with selected variables are presented, the significance of the difference in the correlations of a studied variable with $K^2(X,T)$ and μc is tested and possible relations with the other variables are discussed. The significance of the difference was tested by using Hotelling's method (Walker and Lev, 1953, 258-259).

1. Number of Cases (N)

Table 4 presents the correlations of the three coefficients with the number of cases (the number of students who took the test).

Table 4. Correlations of the Number of Cases with the Three Reliability Coefficients

Test Type	# of Cases		Correlations with			t
	Mean	SD	KR-20	$K^2(X,T)$	μc	
Pretest	163.26	106.91	.10	.09	.10	.08
CET	80.25	58.20	.37**	.05	.33**	2.59**
Posttest	109.40	73.90	.15*	-.01	.22**	2.00*
TOTAL	121.51	91.59	.31**	.17**	.22**	1.18

* Significant at the .05 level.

** Significant at the .01 level.

In general, all three reliability coefficients had positive relationships with the number of cases for the combined total test data. The classical reliability coefficients was mostly highly correlated with the number of cases as expected. Both differences of correlation coefficient of N with KR-20 from those of N with the other reliability coefficients were significant at the .01 ($t = 4.95$) and .05 ($t = 2.23$) level, respectively, whereas, the difference between the latter two coefficients was not statistically significant. The number of students did not show any significant relations with the three reliability coefficients, when the calculations were based on the pretest data.

The KR-20 and μc, however, are significantly related to the number of students involved when the correlations were derived from CET or posttest data. However, the correlation between $K^2(X,T)$ and number of cases was not statistically significant for CET's, or for posttests. Consequently the difference between $r_{N-K^2(X,T)}$ and $_{N-\mu c}$ was significant at the .01 level for the CET case and significant at the .05 level for the posttest case.

Crosstabulation results showed that both KR-20 and μc had distributions of L-shape or extremely positively skewed U-shapes when the number of cases was less than 30. As the number of cases increased, the shape of the KR-20 distributions gradually shifted from the positive to the negatively skewed, while the shape of the μc distributions rapidly shifted from the positive to the negatively skewed.

In short, the above findings imply that Livingston's coefficients are not significantly related to the number of cases, while, the classical internal consistancy coefficient and Harris' index of efficiency are positively correlated with the number of cases. These relationships occured when the tests were administered as posttests (either as immediate or as delayed posttests).

2. Number of Items (K)

It is well known that KR-20 increases as the number of items increases, especially when the items are homogeneous. Livingston's coefficient is expected to have similar relationship with the number of items as KR-20 has because it has KR-20 as a term. Harris' index supposedly does not have any relationship with the number of items. The correlation coefficients of the number of items with the three reliability coefficients are presented in Table 5.

Table 5. Correlations of the Number of Items with the Three Reliability Coefficients

	# of Items		Correlations with			t
Test Type	Mean	SD	KR-20	$K^2(X,T)$	μc	
Pretest	6.22	5.94	.18**	.11	-.11	2.43*
CET	6.72	6.60	.20**	.23**	-.13*	3.17**
Posttest	6.24	6.65	.20**	.19**	-.13*	2.79**
TOTAL	6.38	6.36	.16**	.16**	-.12**	4.72**

* Significant at the .05 level.

** Significant at the .01 level.

As was expected, KR-20 evidenced a moderate positive relation with the number of items for all test types. $K^2(X,T)$ had positive relations with the number of items, even though the correlation coefficient for pretests was not statistically significant. Interestingly, μc had negative correlations with the number of items, and the correlation coefficient for the pretest data was also not statistically significant. Consequently the differences between the correlations of the number of items with $K^2(X,T)$ and with μc were significant at the .05 level for pretests and at the .01 level for the other tests and for the combined total test data. Crosstabulation of K with μc shows that computing μc was adequate when K ≤ 10 or at most 15.

3. Percent Point of Mastery Criterion Score (Pc)

Mastery criterion score for a test was decided on the basis of complexity of the skill and the number of items in the test. Generally, one hundred percent correct was regarded as mastery for a test with less than five items. Lower percent correct were required for tests designed to measure complex skills. Therefore, there is no theoretical basis to expect any relationship between Pc and KR-20, between Pc and $K^2(X,T)$, or between Pc and μc.

Table 6. Correlations of the Percent Point of Mastery Criterion Score with the Three Reliability Coefficients

	Pc		Correlations with			t
Test Type	Mean	SD	KR-20	$K^2(X,T)$	μc	
Pretest	91.28	7.85	-.18**	-.05	.02	.74
CET	91.30	7.51	.00	-.30**	.35**	6.19**
Posttest	92.40	7.56	-.13*	-.19**	.12*	2.75**
TOTAL	91.62	7.66	-.10*	-.18**	.19**	6.20**

* Significant at the .05 level.

** Significant at the .01 level.

Table 6 shows that Pc was negatively correlated with KR-20 and $K^2(X,T)$, and positively correlated with μc. The correlations of Pc with $K^2(X,T)$ and μc for pretests were not statistically significant. The obtained correlations of μc with Pc seem to support the second part of Marshall's (1973) finding that the index is affected by changes in the criterion; the higher the criterion, the higher the index, when the total scores are all close to the number of items. Almost all CET's and most of the posttests were in this case.

4. Percent Point of the Mean ($P\bar{x}$)

When the percent point of the mean approaches an extreme value (0 or 100 percent), the result is a reduction in the test variance, and a concomitant decrease of KR-20. Table 7 shows the decreasing trend well.

Table 7. Correlation Coefficients of the Percent Point of Mean with the Three Reliability Coefficients

	$P\bar{x}$		Correlations with			t
Test Type	Mean	SD	KR-20	$K^2(X,T)$	μc	
Pretest	67.86	22.78	-.34**	-.50**	.04	6.02**
CET	93.51	5.09	-.24**	.17**	-.28**	4.04**
Posttest	90.87	6.72	-.44**	-.20**	-.22**	.21
TOTAL	82.68	19.14	-.44**	-.39**	-.12**	4.96**

*** Significant at the .05 level.**

**** Significant at the .01 level.**

The relationship of $P\bar{x}$ with $K^2(X,T)$ was inconsistent because of the fact that an increase in $P\bar{x}$ effects in two ways two of the most important terms used in determining $K^2(X,T)$ from classical reliability coefficients; namely the standard deviation and $(\mu c-C)^2$. Considering the pretests, where most test means were below the mastery criterion score, an increase in the mean resulted in the reduction of both the test variance and the $(\mu-c)^2$ value. The same reasoning may be applied to the posttest case because the mean of Pc was higher than the mean of $P\bar{x}$ for posttests. For CET's of which the mean of $P\bar{x}$ was higher than the mean of Pc, however, the increase in the mean results in an increase of the $(\mu-c)^2$ term which contributes more than test variance in determining $K^2(X,T)$ for CET where the test variance is usually small.

There were significant negative correlations between $P\bar{x}$ and μc for CET's and for posttests. There were two, nine, and seven 100 percent mastery cases for which the values of μc were zeros in pretests, CET's and posttests, respectively. It is hard to believe, however, that these extreme cases were the sole reasons for the negative correlations for the CET's and the posttests. In this regard, the present results do not agree with Marshall's findings that the index is generally higher as the mean of the test increases for a given criterion level.

5. Difference Between the Mean and the Mastery Criterion Score ($|P\bar{X} - Pc|$ and $|(\bar{X} - C)/SD|$)

As indirectly suggested in the previous discussions of Pc and $P\bar{X}$, the difference between the mean and the mastery criterion score has a close relationship with the magnitude of $K^2(X,T)$. Tables 8 and 9 present the relationships.

Table 8. Correlations of the Difference between the Mean and the Mastery-Criterion Score Expressed in Percentage with the Three Reliability Coefficients

	$\|P\bar{X} - Pc\|$		Correlations with			t
Test Type	Mean	SD	KR-20	$K^2(X,T)$	μc	
Pretest	25.03	22.44	.25**	.48**	-.06	6.43**
CET	7.65	5.09	-.08	.42**	-.32**	7.23**
Posttest	7.45	6.06	.14*	.25**	.09	1.46
TOTAL	14.39	16.99	.31**	.42**	.01	7.50**

* Significant at the .05 level.

** Significant at the .01 level.

Table 9. Correlations of the Difference between the Mean and the Mastery-Criterion Score Expressed in a Standard Score Form with the three Reliability Coefficients

	$\|(\bar{X}-C)/SD\|$		Correlations with			t
Test Type	Mean	SD	KR-20	$K^2(X,T)$	μc	
Pretest	.86	.78	.05	.34**	-.10	4.87**
CET	.83	.97	-.37**	.42**	-.66**	12.87**
Posttest	.55	.70	-.24**	.14*	-.32**	4.29**
TOTAL	.76	.84	-.16**	.32**	-.43**	13.92**

* Significant at the .05 level.

** Significant at the .01 level.

According to Tables 8 and 9, $K^2(X,T)$ was consistantly highly correlated with the difference between the mean and the mastery-criterion score expressed in both percentage and standard score forms for all test types. Obviously, the bigger the discrepency, the larger Livingston's coefficient. The discrepency in percentage form seems more directly related to the magnitude of $K^2(X,T)$ than when it was expressed in standard score form. It is interesting to note that contrary to expectancy, μc was negatively correlated with the discrepency expressed in standard score form. Correlation coefficients were significantly high for the CET and posttest where the test variances were relatively small.

6. Proportion of Mastery Students (Pm)

The relationship between the proportion of mastery students and the three coefficients was investigated separately from that of $P\bar{X}$, even though they were closely correlated (.92 for pretest, .83 for CET, .78 for posttest and .93 for the combined total test data), because Pm has practical significance for decision-making. Obtained correlations are presented in Table 10.

Table 10. Correlations of the Percent of Mastery Students with the Three Reliability Coefficients

	Pm		Correlations with			t
Test Type	Mean	SD	KR-20	$K^2(X,T)$	μc	
Pretest	54.52	26.65	-.22**	-.43**	.21**	7.66**
CET	87.49	11.21	-.19**	.24**	-.40**	6.05**
Posttest	83.97	11.51	-.26**	-.01	-.26**	2.25*
TOTAL	73.54	24.28	-.39**	-.31**	-.15**	3.01**

* Significant at the .05 level.

** Significant at the .01 level.

According to Table 10, KR-20 was significantly negatively correlated with Pm. The results seem reasonable because the increase of Pm might mean the reduction of test variance. In this regard, it does not seem appropriate to compute KR-20 for a criterion-referenced test, especially when it is administered as a CET or as a posttest.

$K^2(X,T)$ did not demonstrate a consistent relationship with Pm. It requires further studies. The μc has a positive correlation for pretest and negative correlations for CET and posttest. When Pm arrives at an extreme value (0 or 100%), μc becomes zero like an ordinary correlation coefficient. There was one Pm = 100 and $\mu c = 0$ case among the 265 pretest cases, 25 among the 208 CET's, and 11 among the 205 postests. Obviously these extreme cases influenced the size of the correlations for the CET and posttest cases. However, one would still not expect to find significant positive correlations for the CET and posttest cases even if these extreme cases were eliminated.

7. Shape of Score Distribution (SSD)

Shape of score distribution is a categorical variable. According to Harris (1972-b), the maximum value of μc^2 is expected to vary along with the shape of score distribution. For symmetric distributions of equal range, a rectangular distribution gives a larger maximum μc^2 than does a normal distribution, and a U-shaped distribution has a larger maximum μc^2 than does a rectangular distribution.

Therefore, value 1 was assigned to one-point distributions, value 2 to a bell-shaped distribution, value 3 to a rectangular or a right-triangle shaped distribution with a gradual slope, value 4 to a J-shaped distribution, and value 5 to a very steep J-shaped distribution with 2 or 3 entry points. Correlations of the categorical variable with the coefficients are presented in Table 11.

Table 11. Correlations of the Shape of Score Distribution with the Three Reliability Coefficients

Test Type	SSD		Correlations With				t
	Mean	SD	KR-20	$K^2(X,T)$	μc	Max μc^2	
Pretest	3.22	1.28	.01	-.15*	.26**	.07	4.47**
CET	3.90	1.04	.36**	.01	.40**	.55**	3.82**
Posttest	4.03	1.10	.43**	.01	.56**	.48**	6.07**
TOTAL	3.68	1.22	.10*	-.14**	.32**	.33**	8.05**

* Significant at the .05 level.
** Significant at the .01 level.

The data in Table 11 seem to support Harris' intuition with one exception; the correlation between SSD and maximum μc^2 is not statistically significant when calculated from the pretest data. The low correlation seems to have resulted from the fact that Max μc^2

had a very small standard deviation.

C. Relations of the Three Reliability Coefficients to $I_{pre-CET}$, $I_{pre-post}$, and $I_{CET-post}$.

Each of the three indices, $I_{pre-CET}$, $I_{pre-post}$ and $I_{CET-post}$, actually represents a compound effect, at least, of the reliability of the two tests used and of the effectiveness of instruction. Therefore, the correlation coefficients shown in Table 12 may be inflated ones.

Table 12. Correlations of the Three Reliability Coefficients with the $I_{pre-CET}$, $I_{pre-post}$, and $I_{CET-post}$ Indices.

Test Type		Index			Correlation with			t
		# of Pairs	Mean	SD	K -20	$K^2(X,T)$	μc	
Pretest	$I_{pre-CET}$	206	84.85	13.56	.10	-.01	.15*	1.57
	$I_{pre-post}$	205	68.86	22.29	-.08	-.13*	.14*	2.55*
CET	$I_{pre-CET}$	206	84.85	13.56	-.18**	.23**	-.41**	6.11**
	$I_{CET-post}$	203	55.46	25.65	-.08	.17**	-.17**	2.96**
Posttest	$I_{pre-post}$	205	68.86	22.29	-.24**	.04	-.26**	2.68**
	$I_{CET-post}$	203	55.46	25.65	-.26**	.03	-.23**	2.37*

* Significant at the .05 level.
** Significant at the .01 level.

Table 12 shows a contrasting tendency between $K^2(X,T)$ and μc for pretest and for CET and posttest. For the pretest data, μc was positively correlated to $I_{pre-CET}$ and $I_{pre-post}$ indices. On the other hand, $K^2(X,T)$ was negatively correlated, though the first correlation coefficient was not statistically significant. However, this tendency was reversed for the CET and posttest data: $K^2(X,T)$ was positively correlated (though the correlation coefficients for the posttest data were not significant), and μc was significantly negatively correlated. More studies seem necessary on the relationship between the test reliability of a CRT and its actual classification ability of students into one of mastery and non-mastery categories.

SUMMARY AND CONCLUDING REMARKS

The present study is a part of the overall study that was designed to find clues to the quetions: (1) what kinds of reliability coefficients are appropriate for various criterion-referenced testing situations, and (2) what are the most appropriate ways of interpreting these coefficients when they are computed.

Livingston's reliability coefficients and Harris' indices of efficiency were computed for 678 criterion-referenced tests in the A through E levels of I.P.I. Mathematics, Edition II. The coefficients were carefully studied and compared with each other and with the classical internal consistency coefficients, KR-20's, in relation to the number of students, number of items, percentage points of the mastery criterion score and the mean, the absolute value of the difference of the mean from the mastery criterion score expressed both in percentage and in standard score form, the standard deviation, the percent of the mastery students, the shape of the score distribution, and the mastery status indices derived from the cross-tabulate tables of students' performance on the pretest and curriculum embedded test (CET), the pretest and posttest, and the CET and posttest.

Generally the means of Harris' indices were larger than those of Livingston's coefficients for all test types (pretest, CET and posttest).

All three reliability coefficients investigated in the present study were higher when a criterion-referenced test was administered as a pretest than when it was used as a CET or as a posttest.

The classical internal consistency coefficient, KR-20,was found to be highly, positively correlated with the standard deviation. The number of cases and the number of items were moderately correlated with KR-20. KR-20 was negatively correlated with the percentage point of the mean.

Livingston's coefficient was highly correlated with the discrepency between the mean and the mastery criterion score. The standard deviation was also highly correlated with Livingston's coefficient for pretest and posttest cases.

When derived from the pretest data, Harris' index showed no significant relation to any variable studied with the exception that it was moderately, positively correlated with the proportion of mastery students and the shape of score distributions. This trend changed when criterion-referenced tests were given either as CET's or as posttests. Harris' index was negatively correlated with the discrepency between the mean and the mastery criterion score, the proportion of mastery students, and interestingly enough with the number of items. It was positively correlated with the number of students who took the test. The shape of the score distribution maintained the same trend as was found with Harris' index based on pretest.

As mentioned before, the present paper is only a report of the descriptive part of the overall study. On the basis of the data presented to date, it would be concluded that Harris' index is relatively stable in regard to all testing situations considered. Livingston's coefficient seems to require different standards for interpretation when it is based on data collected in different testing situations. However, the present author feels that any final conclusions and specific implications for the interpretation of the two reliability coefficients should wait until the following on-going studies are completed; (1) comparisons of the three coefficients in relation to each of the variables mentioned previously, and (2) the analyses of the relative amounts of the contribution each variable made in deciding the size of the reliability coefficients.

REFERENCES

Dixon, W. J. (ed.) *BMD, Biomedical computer program*. Berkeley: Univ. of California Press, 1970.

Edwards, A. L. *Experimental design in psychological research*. 3rd Ed. New York: Holt, 1968.

Hambleton, R. K. and Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Paper presented at the annual meeting of the NCME, Chicago, 1972.

Harris, C. W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. *Journal of Educational Measurement*, 9 (1), 1972-a, 27 - 29.

Harris, C. W. An index of efficiency for fixed-length mastery tests. Paper presented at the annual meeting of the AERA, Chicago, 1972-b.

Livingston, S. A. *The reliability of criterion-referenced measures*. Baltimore; Center for the Study of Social Organization of Schools, the Johns Hopkins University, 1970. (mimeo.)

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

Raju, N. S. A note on Livingston's reliability for criterion-referenced tests. Paper presented at the annual meeting of the NCME, New Orleans, 1973.

Shavelson, R. J., Block, J. H. and Ravitch, M. M. Criterion-referenced testing: Comments on reliability. *Journal of Educational Measurement*, 9 (2), 1972, 133 - 137.

Walker, H. M. and Lev, Joseph. *Statistical inference*. New York: Holt, 1953.