DOCUMENT RESUME

ED 137 359 TM 006 160

AUTHOR Strasler, Gregg M.; Raeth, Peter G.
TITLE An Internal Consistency Estimate for Criterion-Referenced Tests.
PUB DATE [Apr 77]
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, New York, April 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Computer Programs; *Criterion Referenced Tests; Multiple Choice Tests; *Test Reliability

ABSTRACT

The study investigated the feasibility of adapting the coefficient k introduced by Cohen (1960) and elaborated by Swaminathan, Hambleton, and Algina (1974) to an internal consistency estimate for criterion referenced tests in single test administrations. The authors proposed the use of k as an internal consistency estimate by logically dividing criterion referenced tests into two subtests, each tapping mirrored behavioral levels and content areas. Using a computer program developed by the second author, results were tabulated on 93 seventh graders in an experimental study involving a series of multiple-choice tests in the areas of ecology and geometry. (Author/RC)

An Internal Consistency Estimate for Criterion-Referenced Tests

Gregg M. Strasler
University of South Carolina

and

Peter G. Raeth
University of South Carolina

Abstract

The purpose of this study was to investigate the feasibility of adapting the coefficient k introduced by Cohen (1960) and elaborated by Swaminathan, Hambleton, and Algina (1974) to an internal consistency estimate for criterion-referenced tests in single test administrations. The authors proposed the use of k as an internal consistency estimate by logically dividing criterion-referenced tests into two subtests, each tapping mirrored behavioral levels and content areas. Using a computer program developed by the second author, results were tabulated on 93 seventh graders in an experimental study involving a series of multiple-choice tests in the areas of ecology and geometry.

A paper presented at the annual meeting of the National Council on Measurement in Education, New York, 1977.

An Internal Consistency Estimate for Criterion-Referenced Tests

Gregg M. Strasler
and
Peter G. Raeth

University of South Carolina

The purpose of this study was to investigate the feasibility of adapting the coefficient kappa (k) introduced by Cohen (1960) and elaborated by Swaminathan, Hambleton, and Algina (1974) to an internal consistency estimate for criterion-referenced tests in single test administrations.

In an article on the reliability of criterion-referenced tests (Swaminathan et al., 1974), the coefficient kappa (k), an expression for test-retest reliability of criterion-referenced tests, was defined as:

$$k = (P_o - P_e) / (1 - P_e), \tag{1}$$

where P_o, the observed proportion of agreement is given by

$$P_o = \sum_{i=1}^{k} P_{ii} \tag{2}$$

and P_e, the expected proportion of agreement is given by

$$P_e = \sum_{i=1}^{k} P_{i\cdot} \, P_{\cdot i} \tag{3}$$

In these formulas, P_{ii} represents the proportion of examinees placed in the ith mastery state on two test administrations and $P_{i\cdot}$ and $P_{\cdot i}$ represent the proportions of examinees assigned to the mastery state i on the first and second test administrations, respectively. Swaminathan et al. (1974) define k as the proportion of agreement that exists over and above that which can be expected by chance alone.

Swaminathan et al. (1974) define the reliability of a criterion-referenced test as ". . .the measure of agreement between the decisions made in repeated test administrations" (p.264). They further elaborate the need for determining reliability estimates based on "subtest scores" (vis-á-vis objectives) rather than total scores. Although the present writers agree with this conception of test-retest reliability, we propose adapting k as an "internal consistency estimate" for determining the consistency of decision making (i.e., classification of "master" vs. "nonmaster") within a single administration of a criterion-referenced test.[1]

[1]Some persons who are knowledgeable in measurement would disagree with the authors' usage of the term "internal consistency estimate." In the present text, internal consistency estimates include reliability coefficients obtained from single test administrations which are not dependent upon test-taking speed.

Methodology

As stated by Glaser and Nitko (1971), a criterion-referenced test is defined as ". . .one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). In essence, Glaser and Nitko contend that criterion-referenced tests are designed to provide specific information about an individual's performance within a domain of instructionally relevant tasks.

Using this conception of criterion-referenced tests, the first author constructed and administered a series of 20 item, multiple-choice tests (including summative pretests and posttests, and "learning exercises" - formative tests) in a research study involving 93 students in the areas of ecology and geometry. In an effort to establish high content validity, each test was checked by "content experts" (seventh grade teachers) for topic validity as well as process validity (Cureton, 1968). After modifications were made, each test was considered to be a representative sample of both the topics and the cognitive processes of the unit of instruction it represented. Each 20 item test contained six knowledge, seven comprehension, and seven application and analysis items as defined by Bloom's taxonomy (Bloom, 1956). Examples of items categorized by behavior levels are as follows:

(1) Knowledge (concept: food pyramid)

A picture showing that, in an ecosystem, the consumers are fewer than the producers, and the producers fewer than the decomposers is called a:

*a. food pyramid
b. food web
c. food chain
d. niche.

(2) Comprehension (concept: food pyramid)

If we show that the number of herons is smaller than the number of frogs which is smaller in number than the number of crickets, etc., we are showing a picture of a:

a. niche
b. food chain
c. food web
*d. food pyramid

(3) Application (concept: food pyramid)

Choose the statement in which "food pyramid" is used correctly:

*a. The eagles were at the top of the food pyramid.
b. Within the food pyramid, green plants grow.
c. The sides of the food pyramid represent the consumers.
d. Algae are usually the "top dogs" in the food pyramid.

(4) Analysis (concepts: food pyramid, food web)

A food pyramid is different than a food web in that the food pyramid:

a. is a better indicator of what goes on in an ecosystem
*b. describes the numbers of individuals in each population
c. includes producers, consumers, and decomposers
d. shows who eats whom in an ecosystem.

In order to adapt the concept of kappa to measure the internal consistency of a criterion-referenced test, each of the 20 item tests was divided into two, 10 item "subtests." Items within subtests were matched with respect to behavior levels (e.g., knowledge, comprehension, and application and analysis) and content areas. Therefore, the two subtests within each test were approximately the same with respect to difficulty and content covered. It would follow that a student who did well on one subtest would do equally well on the other subtest (and vice-versa) if the test had high internal reliability. If "mastery" is defined as 80% correct, then a master on the first subtest (8, 9, or 10 items correct) should also master the second subtest if the test is internally consistent.

In adapting kappa as an internal consistency estimate, P_{ii} in Equation (2) would represent the proportion of examinees placed in the ith mastery state on two "matched" subtests of a criterion-referenced test. Likewise, $P_{i.}$ and $P_{.i}$ in Equation (3) would represent the proportions of examinees assigned to the mastery state i on the two respective subtests. In essence, kappa may now be interpreted in Equation (1) as the agreement of classification ("mastery" vs. "nonmastery") between subtests after taking into consideration the correction for classifications occurring by chance.

All of the tests were administered to a total of 93 seventh-grade students during twenty-three class days of instruction in the content areas of ecology and geometry. Of the total number of students, 47 students were assigned to a "learning for mastery" (Bloom, 1968) instructional mode and the remaining 46 students served as a control. Although both instructional modes were based on the same objectives and content for each unit of instruction, the learning for mastery students received immediate feedback and corrective procedures for each "learning exercise" (formative test). The learning for mastery students also received additional time in the classroom to correct their mistakes. The control students had neither of the above characteristics.

Results

Through a computer program developed by the second author, the conception of k as an internal consistency estimate was used to analyze the data. Table 1 summarizes some of the output generated by the computer program using k as an internal consistency estimate.

Insert Table 1 about here.

The number or percentages of "masters" or "nonmasters" in the table refer to students who met or did not meet the prespecified criterion (in this case, 70% and 80% correct), respectively, in both subtests of each type of test presented. Therefore, a "master" has the added meaning of achieving "consistency of mastery" on two logically derived subtests as well as meeting a prespecified criterion (70% correct, 80% correct, etc.) on the total test score.

As noted by Swaminathan et al. (1974), the range of k has a lower limit of close to -1 extending to +1 as an upper limit. A negative value of k is, however, indicative of a highly suspect inconsistency in the decision making process. In fact, Millman (1974) points out that, if $k < 0$, the agreement rate would be defined as less than expected by chance. Therefore, as Huynh (1976) suggests, negative values of k should be equated to 0, whereas increasing increments of k in a positive direction should indicate increasing consistency in the decision making process and, hence, increasing reliability. A value of k approximately 0 may be interpreted as what would be expected "by chance" alone. A value where k approximates 0 might also be interpretable in a pretest where no prior instruction has occurred.

It is noteworthy to observe the values of k in Table 1. When a criterion of 80% correct is specified, all of the "learning exercises" (with the exception of the second test in the geometry unit) take on positive values of k after instruction has been received. In the summative pretests (prior to instruction) values of k approximate 0, whereas the values of k for the summative posttests in the ecology and geometry units are 0.438 and 0.580, respectively. The [illegible] overall pattern holds true when the criterion for mastery is set for 70% correct.

Table 2 displays k as a function of criterion for mastery scores for the total group of 93 students as well as for the "learning for mastery" students (N = 47) and the control group (N = 46).

Insert Table 2 about here.

As the criterion for mastery increases, k increases to a limit and then decreases. In general terms, k appears to be "maximal" at the 60% to 80% criterions for mastery scores. These results concur somewhat with Huynh's findings (Huynh, 1976) in which k was maximal at the 65% to 75% criterions for mastery scores for three achievement tests. Huynh (1976) explains this occurrence partly by stating that P_e approximates 1 when the cutoff (criterion for mastery) is too small or too large. Therefore, there is not much room for the "improvement" of the consistency of decisions beyond the chance level.

One other point should be noted from the results obtained in Table 2. The k values for the learning for mastery students (experimental group) appear to be consistently higher than the corresponding k values for

the control group with the exception of the summative pretests in which k approximates 0 for both groups.

In essence, there appears to be a positive relationship between the reliability (i.e., k conceived as an internal consistency estimate) of a logically derived criterion-referenced test and the amount (or quality) of instruction received. If a test is well defined in terms of content covered and behavior levels required, the internal consistency between logically developed subtests may be dependent somewhat on the "meaningfulness" of the instruction received.

Table 3 depicts the means and standard deviations as well as traditional reliability estimates for each of the 10 criterion-referenced tests involved.

Insert Table 3 about here.

The traditional reliability estimates observed were the Kuder-Richardson Formula 20 (KR20) and the Spearman-Brown prophecy formula. The "split-halves" of the Spearman-Brown were identical to the "subtests" used in measuring kappa. As observed in Table 3, there appears to be a high positive relationship between these two traditional estimates of reliability.

Whether or not classical test theory can be applied to a criterion-referenced framework has been a debated issue in recent years. One camp advocates that the concept of variability in test scores is irrelevant with criterion-referenced tests (e.g., Popham and Husek, 1969; Millman and Popham, 1974). The other camp emphasizes that variability has been observed in criterion-referenced testing and is an important concept to be considered (e.g., Woodson, 1974; Haladyna, 1974). The question remains unresolved as to whether classical test measurement (e.g., KR20, Spearman-Brown, etc.) is appropriate for evaluation criterion-referenced tests.

As noted in Table 3, test scores were moderately heterogeneous in the experimental group (learning for mastery students). From a theoretical standpoint, Bloom (1976) indicates that, in such a learning for mastery setting, scores would tend to be more homogeneous. In a mastery learning instructional model, test scores should become higher and less variant in nature.

In an effort to observe what effects would occur in more homogeneous settings, data was simulated (N = 100 cases; test length = 20 items) to

approximate various "stages" of negatively skewed scoring distributions.[1] The kappa coefficient, KR20, and the Spearman-Brown prophecy formula were compared on the following four simulated data sets:

(1) $\overline{X} = 10$, SD = 4

(2) $\overline{X} = 13$, SD = 3

(3) $\overline{X} = 16$, SD = 2

(4) $\overline{X} = 19$, SD = 1.

Data set (1) represents an approximate normal distribution of scores whereas data sets (2), (3), and (4) are negatively skewed with geometrically increasing means ($\overline{X}$'s) and decreasing standard deviations (SD's).

Table 4 denotes the values of kappa (criterion for mastery = 80%), KR20, and Spearman-Brown prophecy formula for the four simulated data sets. When certain properties affecting reliability were held constant (i.e., N = 100 cases; test length = 20 items), all three internal consistency estimates (K, KR20, and Split-Half) were influenced by the decreasing variability in test scores. As variability decreased, so did the internal consistency estimates. In fact, kappa (criterion for mastery = 80%) approximates the value of 0 in data sets (3) and (4). Like the more traditional estimates of internal consistency (KR20, Spearman-Brown), interpretation of kappa becomes suspect when variability in test scores decreases.

Conclusion

The coefficient k appears to be well suited for being used as an internal consistency estimate for criterion-referenced tests in single test administrations. The use of k in test-retest reliability may be too cumbersome a process for teacher-made criterion-referenced tests. There appears to be a need for an internal reliability estimate to indicate the appropriateness of "master" versus "nonmaster" in a single test administration. With the advent of increasing sophistication in criterion-referenced test development, there also appears the need for equating "logical" split-half

[1] In order to preserve the definition of kappa espoused by the authors, the simulated data sets were based on actual item responses made by the 93 students on the third learning exercise in ecosystems. Thus, if two test scores of "18" were required for a simulated data set, two scores of "18" were randomly selected from a pool of students who actually scored "18" on the learning exercise. In the case where there were fewer than three students who actually attained a particular desired score, the next highest score(s) was (were) modified by adding one (or more) randomly selected item(s). Therefore, each of the test scores represented in the simulated data was randomly selected from a pool of at least three or more actual test scores.

reliability by using both behavior levels and content areas as criteria for forming mirrored subtests. Therefore, a "master" (or "nonmaster") has the added meaning of achieving (or failing to meet) a prespecified criterion as well as achieving "consistency of mastery" on two logically derived subtests of a criterion-referenced test.

Unlike other reliability estimates, kappa (k) is concerned with the reliability of classifications, not with the reliability of scores. In a criterion-referenced testing atmosphere, there is a need for consistency in decision-making (e.g., "master" vs. "nonmaster" classification) whether or not variability (in test scores) is present. However, as Swaminathan et al. (1974) have pointed out, kappa is situation specific, and therefore, additional information as criterion for mastery score, test score variability, test length, etc., should be reported along with this index for interpretation.

TABLE 1

Output of Overall Statistics of Kappa Based on 93 Students

Test Type		Criterion for Mastery = 70%				Criterion for Mastery = 80%			
		#M / %M	#NM / %NM	P_o	k	#M / %M	#NM / %NM	P_o	k
Learning Exercises (Ecosystems)	1	6 / 7%	70 / 75%	.817	.306	1 / 1%	82 / 88%	.892	.111
	2	13 / 14%	58 / 62%	.763	.383	3 / 3%	75 / 81%	.839	.195
	3	12 / 13%	48 / 52%	.645	.241	8 / 9%	64 / 69%	.774	.335
Summative Pretest (Ecosystems)		0 / 0%	88 / 95%	.946	-.018	0 / 0%	92 / 99%	.989	.000
Summative Posttest (Ecosystems)		27 / 29%	43 / 46%	.753	.493	14 / 15%	59 / 63%	.785	.438
Learning Exercises (Geometry)	1	9 / 10%	65 / 70%	.796	.391	3 / 3%	77 / 83%	.860	.250
	2	1 / 1%	80 / 86%	.871	.079	0 / 0%	91 / 98%	.978	-.011
	3	6 / 7%	69 / 74%	.806	.285	2 / 2%	82 / 88%	.903	.259
Summative Pretest (Geometry)		1 / 1%	86 / 93%	.935	.236	0 / 0%	91 / 98%	.978	.000
Summative Posttest (Geometry)		12 / 13%	62 / 67%	.796	.428	9 / 10%	74 / 80%	.892	.580

NOTE:

#M = number of masters in both subtests

%M = percentage of masters in both subtests

#NM = number of nonmasters in both subtests

%NM = percentage of nonmasters in both subtests

P_o = the observed proportion of agreement of masters and nonmasters in both subtests

k = proportion of agreement that exists over and above that which can be expected by chance alone

TABLE 2

Kappa As A Function of Criterion For Mastery Scores

Test Type		Total (N=93) Criterion For Mastery 50%	70%	80%	90%	100%	Experimental (N=47) Criterion For Mastery 60%	70%	80%	90%	100%	Control (N=46) Criterion For Mastery 60%	70%	80%	90%	100%
Learning	1	.083	.306	.111	.000	.000	.085	.290	.109	.000	.000	.030	.238	-.022	ONE	ONE
Exercises	2	.433	.383	.195	.000	ONE	.402	.382	.178	.000	ONE	.370	.107	.000	ONE	ONE
(Ecosystems)	3	.349	.241	.335	.069	.000	.368	.222	.287	.075	.000	.280	.143	.335	.000	.000
ξ Pre (Ecosystems)		-.063	-.018	.000	.000	ONE	-.079	-.029	.000	.000	ONE	-.038	.000	ONE	ONE	ONE
ξ Post (Ecosystems)		.511	.493	.438	.294	.000	.599	.421	.386	.301	.000	.249	.388	.340	-.022	ONE
Learning	1	.402	.391	.250	-.022	ONE	.439	.424	.225	-.044	ONE	.258	-.038	.000	ONE	ONE
Exercises	2	.139	.079	-.011	ONE	ONE	.183	.141	-.022	ONE	ONE	-.073	-.036	ONE	ONE	ONE
(Geometry)	3	.256	.285	.259	.258	.000	.276	.217	.332	.368	.000	.033	.287	-.030	-.022	ONE
ξ Pre (Geometry)		.179	.236	.000	.000	ONE	.208	.221	.000	.000	ONE	.000	ONE	ONE	ONE	ONE
ξ Post (Geometry)		.408	.428	.580	.239	.000	.395	.309	.533	.187	.000	.045	-.030	.000	ONE	ONE

NOTE:

Experimental = learning for mastery students

Criterion for Mastery = percent of items correct in both subtests

ONE = the case where no one has mastered either subtest

TABLE 3

Means, Standard Deviations, and Reliability Estimates

Test Type		Total (N=93)			Experimental (N=47)			Control (N=46)		
		Mean (SD)	KR 20	Split-Half	Mean (SD)	KR 20	Split-Half	Mean (SD)	KR 20	Split-Hal
Learning	1	9.02 (3.20)	.605	.590	9.49 (3.67)	.697	.621	8.54 (2.61)	.413	.501
Exercises	2	9.91 (3.43)	.673	.698	11.02 (3.56)	.712	.707	8.78 (2.91)	.533	.614
(Ecosystems)	3	10.69 (3.86)	.731	.774	11.79 (4.03)	.772	.801	9.57 (3.37)	.629	.685
Σ Pre (Ecosystems)		6.69 (2.35)	.270	.146	6.81 (2.59)	.403	.347	6.57 (2.09)	.079	-.163
Σ Post (Ecosystems)		11.38 (4.05)	.775	.808	13.28 (3.71)	.764	.804	9.44 (3.46)	.661	.728
Learning	1	9.33 (3.69)	.719	.763	10.57 (3.93)	.767	.803	8.07 (2.97)	.548	.609
Exercises	2	8.02 (2.44)	.391	.376	9.00 (2.49)	.409	.500	7.02 (1.94)	.083	-.122
(Geometry)	3	9.44 (3.16)	.597	.560	10.62 (3.17)	.602	.564	8.24 (2.69)	.450	.377
Σ Pre (Geometry)		5.62 (2.72)	.544	.544	6.66 (2.97)	.611	.586	4.57 (1.95)	.149	.191
Σ Post (Geometry)		9.46 (4.16)	.782	.765	11.66 (4.18)	.794	.740	7.22 (2.71)	.484	.540

NOTE:

Experimental = learning for mastery students

SD = standard deviation

KR 20 = Kuder-Richardson Formula 20

Split-Half = Spearman-Brown Prophecy Formula

TABLE 4

Internal Reliability Estimates for Four Simulated Data Sets (N = 100)

Data Set	Mean (SD)	Skewness	Kurtosis	K	KR20	Split-Half
1	10.00 (4.05)	0	-.331	.328	.753	.777
2	13.02 (2.97)	-.503	-.035	.126	.575	.644
3	15.98 (1.97)	-.352	-.210	.083	.334	.425
4	18.98 (1.08)	-1.162	1.255	.000	.337	.507

NOTE:

SD = standard deviation

K = kappa coefficient (criterion for mastery = 80%)

KR20 = Kuder-Richardson Formula 20

Split-Half = Spearman-Brown Prophecy Formula

References

Bloom, B. (Ed.) Taxonomy of educational objectives. The cognitive domain. New York: David McKay Company, Inc., 1956.

Bloom, B. Learning for mastery. Evaluation Comment, 1968, 1, Center for the Study of Evaluation, University of California, Los Angeles.

Bloom, B. Human characteristics and school learning. New York: McGraw-Hill, 1976.

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

Cureton, E. Measurement theory. In R. L. Ebel (Ed.), Encyclopedia of educational research. New York: The Macmillan Company, pp. 785-804, 1969.

Glaser, R., and Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement. Washington: American Council on Education, 1971, pp. 625-670.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 93-99.

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

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: current practices. Berkeley, California: McCutchan Publishers, 1974.

Millman, J., and Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 1974, 11, 137-138.

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

Swaminathan, H., Hambleton, R. K., and Algina, J. A. Reliability of criterion-referenced tests: a decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Woodson, M. I. C. E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64.