DOCUMENT RESUME

ED 126 154 TM 005 411

AUTHOR Epstein, Kenneth I.; Knerr, Claramae S.
TITLE Criterion-Referenced Test Interpretations of "Classical" Measurement Theory.
PUB DATE [Apr 76]
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Item Analysis; Item Sampling; Performance Tests; Simulation; *Statistical Analysis; *Test Interpretation; Test Reliability
IDENTIFIERS *Test Theory

ABSTRACT

The literature on criterion referenced testing is full of discussions concerning whether classical measurement techniques are appropriate, whether variance is necessary, whether new indices of reliability are needed, and the like. What appears to be lacking, however, is a clear and simple discussion of why the problems occur. This paper suggests that many of the results obtained when classical techniques are applied to criterion referenced tests, particularly in the context of mastery learning, are perfectly reasonable, interpetable, and should be expected. (Author)

Criterion-Referenced Test Interpretations of "Classical" Measurement Theory

Kenneth I. Epstein
Claramae S. Knerr

Army Research Institute for the
Behavioral and Social Sciences

A paper presented at the American Educational Research Association annual meeting, San Francisco, California April, 1976

Criterion-Referenced Test Interpretations of "Classical" Measurement Theory

Kenneth I. Epstein
Claramae S. Knerr

Army Research Institute for the
Behavorial and Social Sciences

The literature on criterion-referenced testing is full of discussions concerning whether "classical" measurement techniques are appropriate, whether variance is necessary, whether new indices of reliability are needed, and the like (see, for example, Woodson, 1974, a,b.; Millman and Popham, 1974; Haladyna, 1974). What appears to be lacking, however, is a clear and simple discussion of why the problems occur. This paper suggests that many of the results obtained when "classical" techniques are applied to criterion-referenced tests, particularly in the context of mastery learning, are perfectly reasonable, interpretable, and should be expected.

Consider, for example, Nunnally's (Nunnally, 1967) discussion of the domain-sampling model. The model assumes that any particular measure is composed of "a random sample of items from a hypothetical domain of items (p. 175)." The definition of "true score" is the score that would be obtained if all items in the domain were included in the measure. The only other assumption required for the development of the model is that all items contain an equal amount of the "common

The views expressed in this paper are those of the authors and do not imply indorsement by the U.S. Army.

core", the skill or attribute that is being measured. Statistically, this implies that the average correlation of each item with all the others is the same for all items. Notice that this does not imply that all the inter-item correlations are the same. This description seems to fit very nicely with what is required for criterion-referenced tests. In fact, it is almost the same as the situation Millman (Millman, 1973) described in his review article on domain-referenced measures.

It is possible to derive Cronbach's (1951) coefficient alpha from the domain-sampling model without making any further assumptions about the nature of the domain. As Cureton (1958) pointed out, the only required assumption is that for any given k-item test "there is at least one possible division of the k-items into the two half-tests, x_a and x_b, such that these two half-tests are equally reliable and equally variable (p. 725)." Since, in general, the particular partition of items that meets this requirement is not known it is also necessary to assume "that the mean within-half-test item covariances are not only equal to each other but are equal also to the mean between-half-test item covariances (p. 726)". However, these assumptions are merely a restatement in terms of the covariances of the basic assumption of the domain-sampling model that the average correlation of each item with all the others is the same for all items. If this is the case, then the question is, Why doesn't the model work with criterion-referenced tests? We maintain that the

model does work. The problems occur in interpreting the results. If one considers the results of applying classical techniques in terms of the nature of the items and the people being tested, they are exactly what would be expected.

To illustrate the need for careful consideration of the data source before statistical results are interpreted, two testing situations are described. These examples come from a purely military context (in fact, they involve tank gunnery skills), but the tests clearly fit the requirements of criterion-referenced tests. Therefore, the tests themselves and the results of the analyses reported here should be thought of in the general context of criterion-referenced testing and not in the restricted context of military testing problems.

The data come from two separate studies, each designed to investigate different aspects of the use of simulation devices to train tank gunners. The first study (Powers, et al, 1975) investigated the contribution of live firing, as a component of the training program, to gunnery proficiency. Trainees were divided into four training groups and were allowed twenty-four training "rounds" before taking the criterion test. The proportions of live fire and laser simulated fire included during the training differentiated the training groups. Although the four groups received different types of training, the most critical aspect of the training, the number of practice rounds,

was constant for each group. Therefore, it would be reasonable to expect that the relative performances of the training groups would be similar. In fact, the study failed to show differences in performance on the criterion test as a function of the training program.

The second study (Rose, et al, in press) was designed to compare the relative effectiveness of two tank gunnery training devices. Seven training groups were included in the study. Three groups were assigned to each training device and trained to proficiency levels on the device of 30%, 50%, and 70% hits. The seventh group received no training. Since the groups received different amounts of training and achieved different proficiency levels in training, it would seem reasonable to expect different levels of performance on the criterion test. In fact, the study did reveal that some of the variance in performance on the criterion test could be attributed to differential levels of proficiency in training.

The criterion test in each study required gunnery trainees to demonstrate their ability to hit moving targets with the main gun of the M60A1 tank. In the first study, each gunner fired eight rounds, each of which was scored hit or miss. In the second study, the test consisted of twelve rounds. In the first study, the range to the target was 1400 meters during its movement from left to right, and 1200 meters during its movement from right to left. The corresponding ranges for the second study were 750 meters and 700 meters. With the above exceptions of range and direction of travel, each test trial

was identical. Both tests seem to fit Nunnally's domain sampling model and Millman's requirement that all items come from the same domain. Yet the analyses of the test scores yield vastly different and seemingly contradictory results.

Summary statistics for the criterion tests are provided in Tables 1 and 2 for the first and second study, respectively. Figure 1 compares the frequency distributions. The values for KR-20 and the average interitem correlation are particularly perplexing. The first study's results indicate a KR-20 of .0048 and an average interitem correlation of .0027. Further, of the 28 intercorrelations obtained from the item intercorrelation matrix, only 3 are significantly different from zero at the .05 level of significance. The second study yielded very different results. The value of KR-20 was .7136, and the average interitem correlation was .1716. The correlation matrix for the second study showed 41 of the 66 possible intercorrelations significantly different from zero at the .05 level. Why should two such apparently similar tests show such different results? Perhaps more importantly, how does a test constructor interpret these results and how can he use them to design better tests?

TEST SCORE	FREQUENCY	ITEM	DIFFICULTY (P)
0	1	1	.345
1	10	2	.418
2	11	3	.236
3	18	4	.345
4	8	5	.400
5	6	6	.418
6	1	7	.273
7	0	8	.364
8	0		
	55		

Variance: 1.797
KR-20: 0.0048
Average interitem correlation: 0.0027

Table 1. Summary statistics for the criterion test for Study 1

TEST SCORE	FREQUENCY	ITEM	DIFFICULTY (P)
0	2	1	.429
1	2	2	.487
2	10	3	.526
3	12	4	.474
4	13	5	.500
5	16	6	.558
6	20	7	.545
7	16	8	.662
8	19	9	.578
9	14	10	.636
10	14	11	.604
11	11	12	.630
12	5		
	154		

Variance: 8.402
KR-20: 0.7136
Average interitem correlation: 0.1716

Table 2. Summary statistics for the criterion test for Study 2

Measurement requires that there be sampling among people and among items. Typically, the sampling problem with regard to people is ignored during test development or item writing before validation. One simply assumes that sufficient people are available and that there is sufficient variability in their abilities to allow for item characteristics to be studied. Essentially all of the test developer's time is spent in sampling items and assuring himself that they are good ones. However, in a criterion-referenced testing/mastery learning context both sampling problems must be considered. What are the implications for measurement theory if the examinee population has very little variability in ability? Hambleton and Traub (1973) in their article on latent trait models provide an answer to this question in the discussion of "local independence": "... in an infinite subpopulation of examinees, all of whom are at the same ability level, scores on one test item will be statistically independent of scores on another (if the assumption of local independence holds). It will be recognized that the assumption of local independence does <u>not</u> imply that test items are uncorrelated over the total group examinees. Correlations between items measuring the same ability will, in general, exist whenever the examinees responding to the items differ on the underlying ability measured by the test (pp. 195-196)." The data from the first study illustrate the results of using a good, content valid criterion-referenced test with a highly homogeneous examinee population. The second study illustrates the results of using a similar test with a

more traditional relatively heterogeneous examinee population. The two sets of data behave in precisely the manner described by Hambleton and Traub. In a mastery learning context a homogeneous examinee population is expected or, produced by the training. Perhaps if this type of situation prevails, a criterion-referenced test will produce low values of KR-20 and zero interitem correlations. Certainly, such results should not be cause for alarm.

One final interesting feature of these data should be mentioned. If the series of test trails is considered a series of independent Bernoulli trials, then the average proportion of hits is an unbiased estimate of the group ability. For the first study the average proportion of hits was .35. For the second study the average proportion of hits was .55. One can compare the theoretical characteristics of a series of Bernoulli trails with p=.35 and p=.55 to the observed data. In the case of the first study the Kolmogorov-Smirov one-sample test (Siegel, 1956, p.47) indicates that the probability of obtaining the scores observed under the null hypothesis that the score distribution is Bernoulli with p=.35 is greater than .20. In other words, it seems reasonable to explain the observed test results in terms of a highly homogeneous group of individuals, each of whom has a .35 chance of hitting the target.

For the second study the null hypothesis is that the score distribution is Bernoulli with p=.55. The Kolmogorov-Smirnov test indicates that the probability of obtaining the observed data under

the null hypothesis is less than .01. Hence, the assumption that the second study dealt with a heterogeneous examinee population seems to be reasonable. In fact, the data from the second study fit a negative hypergeometric distribution with parameters n=12, a=2.732, and b=13.217 with a probability greater than .20 using the Kolmogorov-Smirnov test, and provide a good example of an application of the binomial error model discussed in Lord and Novick (1968, pp. 508-529). Thus, these tests support the conclusions regarding the relationship of the examinee group, interitem correlations, and local independence in the Hambleton and Traub paper.

The purpose of this paper is not to advocate a particular procedure for evaluating criterion-referenced tests. Rather, it is to remind the practitioner that more than statistics and measurement theory are required in order to interpret test results meaningfully, and to provide two examples which illustrate the importance of considering the entire testing situation in making inferences about a particular test. While the areas of the relationship between examinee ability and test reliability and the implications of item independence have been addressed in the literature, it appears that the widespread application of criterion-referenced testing requires that these subjects be reexplored. Particularly, the statistical properties of restricted distributions, such as the abilities of students in a mastery learning program, should be reexamined to determine their applicability in interpreting criterion-referenced test results. Perhaps by explicitly defining what is known, the directions for further research will become more clear.

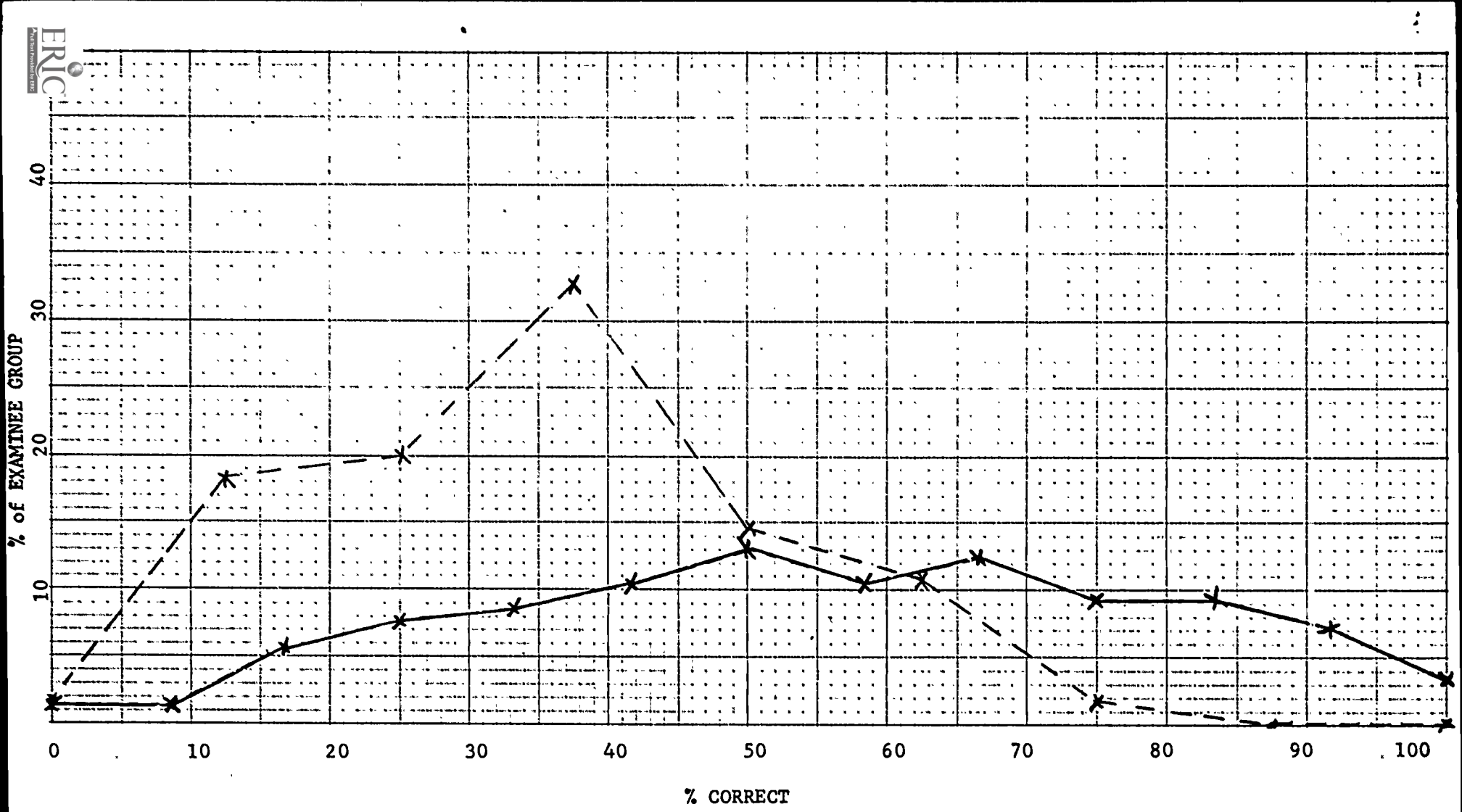

Figure 1: Observed score distributions for Study 1 – – – – and Study 2 ————.

References

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cureton, E.E. The definition and estimation of test reliability. Educational and Psychological Measurement, 1958, 18, 715-738.

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

Hambleton, R.K. and Traub, R.E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.

Lord, F.M. and Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley Publishing Co., 1968.

Millman, J. Passing scores and test lengths for domain referenced tests. Review of Educational Research, 1973, 43, 205-216.

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.

Nunnally, J.C. Psychometric theory. New York: McGraw-Hill Book Co., 1967.

Powers, T.R., McCluskey, M.R., Haggard, D.F., Boycan, G.G., and Steinheiser, F. Jr. Determination of the contribution of live firing to weapons proficiency Alexandria, Va.: Human Resources Research Organization, Final Report FR-CD(C) 75-1, March, 1975.

Rose, A.M., Wheaton, G.R., Leonard, A.L., Fingerman, P.W. and Boycan, G.G. Evaluation of two tank gunnery trainers. Arlington, Va: U.S. Army Research Institute for the Behavioral and Social Sciences, in press.

Siegel, S. Nonparametric statistics for the Behavioral sciences. New York: McGraw-Hill Book Company, 1956.

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

Woodson, M.I.C.E. The issue of item and test variance for criterion referenced tests: A reply. Journal of Educational Measurement, 1974, 11, 139-140.