DOCUMENT RESUME

ED 050 139 TM 000 526

AUTHOR Hsu, Tse-Chi

TITLE Empirical Data on Criterion-Referenced Tests.
INSTITUTION Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY Department of Health, Education, and Welfare, Washington, D.C. Office of the Commissioner of Education.
PUB DATE Dec 71
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS *Behavioral Objectives, Correlation, *Criterion Referenced Tests, *Individualized Instruction, *Item Analysis, Statistical Analysis, Test Construction, *Test Reliability
IDENTIFIERS Individually Prescribed Instruction Program, IPI

ABSTRACT

A good criterion-referenced test item is defined as the one which allows the individual to answer correctly if he masters the criterion behavior represented by the item and answer incorrectly if he actually does not master it. Therefore, a good discriminating item for criterion-referenced tests is the one which has a larger proportion of correct responses in the mastery group and a smaller proportion of correct responses in the non-mastery group. Based on these considerations, the difference in proportions of correct responses in mastery and non-mastery groups and the phi coefficient are proposed as discrimination indices for criterion-referenced test items. These two indices were compared empirically with the point biserial correlation of items and test scores in three different situations: heterogeneous sample with a symmetrical distribution of scores, a homogeneous sample with a skewed distribution, and varying item difficulties. Results indicate the indices are highly correlated in most cases. Implications of these comparisons are noted. A possible approach for criterion-referenced test reliability is also discussed. (Author/IG)

Empirical Data on Criterion-Referenced Tests

Tse-Chi Hsu
Learning Research and Development Center
University of Pittsburgh

A paper presented at the Annual Meeting of the American Educational Research Association, New York City, New York, February 4-7, 1971.

The preparation of this paper was supported by the Learning Research and Development Center supported as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare.

Empirical Data on Criterion-Referenced Tests

Although criterion-referenced measurement is not a brand new invention, its recent marriage with individualized instruction and instructional technology attracts some interest in the area of measurement. Unfortunately, not too many people can agree on exactly what criterion-referenced measurement is. Therefore, I do not expect you to agree entirely with my version of criterion-referenced tests. In this paper I will describe some of our experiences in the analysis of criterion-referenced test items for the Individually Prescribed Instruction (IPI) program at the University of Pittsburgh.

The version of criterion-referenced tests which will be used in this discussion is structured as shown in the sample in Table 1. Let us assume there are four behavioral objectives (or classes of behavior). The number of items and mastery levels are identical for each objective in this example, but this is a coincidence. In actual practice, the number of items per objective and the mastery level for each objective may vary according to the nature of the objective. A mastery level is defined as the cut-off score which is used to declare a student a master or non-master for each criterion behavior. It is not unusual for a test to consist of only a single objective, especially if the test is going to be used in instruction. However, if more than one objective is included in a test, items for each objective can be grouped together as a subtest. Test scores referred to in this paper are subtest scores, that is, a separate score for each objective.

Item Selection Procedures

Since the criterion-referenced test is used to distinguish mastery or non-mastery of certain criterion behaviors, rather than to differentiate individuals in a group, several new item discrimination indices have been proposed. Cox and Vargas (1966) computed the percentage of students who passed an item on the posttest minus the percentage of those who passed an item on the pretest. Popham (1970) used chi-square to contrast the pre- and post-instruction relation of each item with hypothetical frequencies based on the median value of each subtest. Rahmlow, Matthews, and Jung (1970) suggested the combined use of the difficulty level and the instruction gain scores in analyzing criterion-referenced test items. Although these procedures are different, they have one thing in common--the use of instruction as a basis of discrimination.

The difficulty of using instruction as a basis of discrimination is that instruction is not necessarily equal to learning. Poor instruction may have negative effects on item statistics. In terms of time and money, the tryout procedure of using instruction as a necessary component is not very economical. If the same test is used as both pre- and posttest, there is the question of whether the student just learned the specific items on the test or the general class of behaviors which the items sample.

So far, there are no adequate statistical indices one may use to select items for criterion-referenced tests. I will not attempt to present any item selection data using new indices proposed for criterion-referenced tests. Instead, I will reexamine the meaning of criterion-referenced tests and see how classical item selection procedures may be applied. If the test is going to be used to provide "explicit information as to what the individual can or cannot do" (Glaser, 1963, p. 520), then a good criterion-referenced test item does not only discriminate pre-

and post "learning." It is also the function of the item to allow the individual to answer correctly if he masters the criterion behavior represented by the item and answer incorrectly if he actually does not master it, regardless of whether the test is administered before or after formal instruction.

Empirically, the person who masters a criterion behavior is the one who was declared to have mastery on the test. Therefore, in the tryout of items, we may use the predetermined cut-off score for each behavior to declare a mastery group and a non-mastery group. Then, for each item, we may obtain the proportion of subjects who answered correctly in the mastery group and the proportion of subjects who answered correctly in the non-mastery group. The difference of these two proportions (D_p) is a meaningful discrimination index for criterion-referenced test items. An item which has a larger proportion of correct responses in the non-mastery group certainly is not a good representation of its corresponding behavior.

Another way to compute this type of discrimination is to use the phi (ϕ) coefficient. By using right (1) or wrong (0) response with the mastery (1) or non-mastery (0) of the subject, the ϕ coefficient can be obtained easily. Although the ϕ coefficient lacks invariance properties, it is probably a better index than that of Tetrachoric correlation, since Tetrachoric correlation is very difficult to compute and the bivariate normal distribution should be assumed.

This phi (ϕ) index is a minor modification of classical item total score correlations to fit the idea of criterion-referenced tests. The usefulness of the index deserves careful examination. First, we will discuss the limitations of the this index. Then, a comparison of point biserial correlation with the phi (ϕ) coefficient and difference in

proportions (D_p) of correct response in mastery and non-mastery groups will be presented.

The ϕ coefficient is ambiguous and cannot be solved when: (a) the item is answered correctly or incorrectly by everyone, or (b) all subjects are declared as mastery or non-mastery. In these situations, the difference in proportions between mastery or non-mastery groups (D_p) or the point biserial correlation coefficient (r_{pbi}) probably makes more sense.

Empirical comparisons of r_{pbi}, ϕ, and D_p were designed to investigate:[1]

1) the relationship among r_{pbi}, ϕ, and D_p within a sample when (a) the sample consists of subjects with wide variety of abilities and (b) the sample consists of subjects with homogeneous ability skewed to one side;

2) the consistency of r_{pbi}, ϕ, and D_p from one sample to another when (a) the samples have similar test score distributions, and (b) the samples have different test score distributions; and

3) the relationship among r_{pbi}, ϕ, and D_p when items vary in difficulty.

Two similar studies were performed. In the first study, the pre- and posttests of IPI math, D-subtraction[2] were used. The pre- and posttests are equivalent but not identical. These tests were developed from the objectives presented in Table 1, five items for each objective. Data of IPI students taking these tests were obtained and compared to those of

---

[1]The author is indebted to Miss Betty Boston for her assistance in administering the tests.

[2]IPI Mathematics, Developmental Edition, Appleton-Century-Crofts, 1967. One of the objectives in D-subtraction was not used because it is a timed test.

non-IPI students who took the same pre- and posttests. Descriptive statistics of the results are shown in Table 2. In the pretest, the score distributions for IPI students and non-IPI students are not too far apart. However, since the posttest for the IPI group was given after instruction, the scores of the IPI students on the posttest are far more homogeneously scattered to the right-hand side. For each item, r_{pbi}, ϕ, and D_p were computed in each sample. Then, the intercorrelations of these indices were calculated by the Pearson-product moment correlation. These data are presented in Tables 3 and 4.

In the second study, another two forms of the D-subtraction test for the same objectives were constructed, with four items for each objective. Form A was administered to students in grades 3 and 4. Form B was administered to students in grades 2 and 3 in different schools. As shown from Table 5, the variations for the two groups that took Form A are not substantial. The second grade group in Form B, Table 6, are highly skewed to the left. The intercorrelations of r_{pbi}, ϕ, and D_p for these two tests are given in Tables 7 and 8.

From Tables 3, 4, 7, and 8, one may observe that when the sample consists of subjects with a wide variety of abilities and with more symmetrical distribution of test scores, r_{pbi}, ϕ, and D_p are all highly correlated to each other. When the sample consists of homogeneous subjects and skewed to either left or right, the correlations between r_{pbi} and ϕ and between r_{pbi} and D_p, though all significant at the 5 per cent level, are considerably lower. This trend is shown in the IPI group, Table 4, and again in Grade 2, Table 8.

The consistency of r_{pbi}, ϕ, and D_p from one sample to another can be judged from Tables 3, 4, 7, and 8 too. The correlations between two samples for r_{pbi}, ϕ, and D_p are all very high in Tables 3 and 7 and

rather low in Tables 4 and 8. Evidently when samples having similar test score distribution, r_{pbi}, ϕ, and D_p are all relatively consistent from one sample to another. However, when samples have differently shaped score distributions, these discrimination statistics cannot be generalized from one sample to another. Thus, a highly discriminate item for a group with a wide variety of abilities is not necessarily a highly discriminate item for a selected group. Therefore, an identical item may not measure the same type of performance in two different groups. In other words, if we attempt to tryout test items on a group which has a wide variety of abilities in order to apply these discrimination indices, and use this information to select high discrimination items for a second highly selected group, such as the IPI group in Table 4, we may not be choosing the appropriate items.

What is the effect of item difficulty on the correlations among r_{pbi}, ϕ, and D_p? Items were grouped into three categories according to the index of difficulty: high (.7 and higher), middle (.4-.7), and low below .4). Four samples of correlations for these three indices were obtained. Table 9 shows that, as one may expect, the correlations are consistently higher when items are in the middle difficulty.

These empirical data show that ϕ and D_p are consistent with r_{pbi} in most cases. To use ϕ or D_p in the selection of items for norm-referenced tests may not be justifiable because they tend to lose some information. For criterion-referenced tests, if our definition of good criterion-referenced test items is reasonable, then ϕ and D_p are simple ways of detecting poorly discriminated items. It should be emphasized that any good index is useful only to the extent that it helps in differentiating items according to certain characteristics. It should not be used as the only basis for item elimination. Ultimately a human judgement is required to decide whether

an item should be revised or eliminated by considering the statistical properties of items and test scores in light of the purpose and the nature of the test.

Reliability of Criterion-Referenced Tests

It is generally known that the reliability coefficients computed from the traditional reliability formulas are affected by the heterogeneity of test scores. Since the criterion-referenced tests are not designed to produce variability in test scores, they obviously cannot avoid the problem of the homogeneity of test scores. Also, too much emphasis on the homogeneity of items is not desirable either because it may reduce the validity of the test.

To apply classical reliability formulas for a criterion-referenced test by disregarding different behavioral objectives within a test is evidently undesirable. Items for different objectives should be treated separately. If we can compromise the homogeneity of items within an objective with the heterogeneity across the objectives, the homogeneity of items is not necessarily a bad property for criterion-referenced measurement. In other words, if we can increase traditional reliability for items within one objective but decrease the homogeneity across the objectives, we may be able to increase both reliability and validity at the same time. However, this concept needs further exploration and empirical investigation.

To deal with the homogeneous subjects' problem, Livingston (1970) suggested an alternate type of reliability coefficient. He used the cut-off score that defines mastery, instead of the mean, to redefine deviation for criterion-referenced testing. The reliability coefficient computed by this method is at least as large as the norm-referenced reliability. The greater the difference between the mean and the cut-off score, the greater this reliability coefficient will be. Thus, changing the cut-off score will change this reliability coefficient.

Let us refer back to Table 6. Table 6 shows two groups with differences in score distributions. The reliability coefficients computed from KR20 become very low as the items become very difficult for the second grade group. Using a criterion of 75 per cent, the corresponding Livingston coefficient (r_c^2) can be increased considerably. However, one may still question whether a test is going to be more reliable if the difference between the mean and the cut-off score is maximized.

Summary

Criterion-referenced measurement represents an attempt to measure and to interpret human behaviors more meaningfully. In view of recent developments in individualized instruction and instructional technology, the traditional approach of comparing a student's performance with his peers is not enough. This is especially true if the test results are going to be used to make a decision about further instruction. To be able to judge what a student can and cannot do, items that yield a better prediction of what a person can and cannot do should be used and grouped together for the convenience of interpretation and item analysis. A mastery level should be determined for each criterion behavior (or objective) rather than judging a group of objectives as a whole. The mastery level will not necessarily be the same for every objective in a test.

A criterion-referenced test designed in this manner can be greatly facilitated by the item analysis procedures and the application of classical reliability theory. We examined the possibility of using the phi (ϕ) coefficient and the difference in proportions of correct responses between mastery and non-mastery groups (D_p) as discrimination indices for criterion-referenced test items. These two indices were also compared empirically

with the point biserial correlation (r_{pbi}) between item and test score under three different situations. Results showed that these indices are highly correlated in most cases. However, because of the inconsistency of these indices from a group with wide variety of abilities to a highly selected group, the items selected for one group may not be measuring the same kind of performance in a second more homogeneous group. Therefore, the procedure of trying out test items in a group with wide variety of abilities in order to apply these indices in selecting items for criterion-referenced tests is not recommended.

References

Cox, R. C. and Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Illinois, February, 1966.

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

Livingston, S. A. The reliability of criterion-referenced measures. Report No. 73. The Center for the Study of Social Organization of Schools, Johns Hopkins University, 1970.

Popham, W. J. Indices of adequacy for criterion-referenced test items. Paper presented at the annual meeting of NCME/AERA, Minneapolis, Minnesota, March, 1970.

Rahmlow, H. F., Matthews, J. J., and Jung, S. M. An empirical investigation of item analysis in criterion-referenced tests. Paper presented at the annual meeting of NCME/AERA, Minneapolis, Minnesota, March, 1970.

Table 1

A Sample Structure of a Criterion-Referenced Test

	Objective*	No. of Items	Mastery Level
I.	Does subtraction without borrowing for numbers with three or more digits.	5	80%
II.	Subtracts with borrowing from tens place using two-digit numbers.	5	80%
III.	Subtracts with borrowing from tens or hundreds place with three-digit numbers.	5	80%
IV.	Subtracts with borrowing from tens and hundreds place with three-digit numbers.	5	80%

*The objectives are taken from IPI math continuum, 1968-69 (working copy). Learning Research and Development Center, University of Pittsburgh.

Table 2

Descriptive Statistics of IPI and Non-IPI Subjects
Using the Same Pre- and Posttests of IPI Math,
D-Subtraction

			IPi (N=77)			Non-IPI (N=78)		
	Objective I.D.	No. of Item	Mean	St. Dev.	KR20	Mean	St. Dev.	KR20
Pre	I	5	4.45	1.09	.75	4.51	1.05	.77
	II	5	2.79	2.09	.90	1.99	2.33	.97
	III	5	2.42	2.32	.96	1.78	2.27	.97
	IV	5	1.90	2.11	.93	1.36	1.96	.93
			(N=59)			(N=78)		
Post	I	5	4.69	.73	.59	4.58	.89	.65
	II	5	4.47	1.19	.83	1.69	2.28	.98
	III	5	4.00	1.36	.72	1.62	2.17	.96
	IV	5	3.64	1.50	.72	1.10	1.77	.91

Table 3

Intercorrelations of r_{pbi}, ϕ, D_p from IPI and Non-IPI Subjects
Using the Same Pretest of IPI Math,
D-Subtraction (N = 20 items)

	Pretest (IPI)			Pretest (Non-IPI)		
	1. r_{pbi}	2. ϕ	3. D_p	4. r_{pbi}	5. ϕ	6. D_p
1		.77	.81	.85	.76	.68
2			.82	.73	.67	.53
3				.81	.76	.68
4					.78	.75
5						.88

Note: 5% = .444, 1% = .561

Table 4

Intercorrelations of r_{pbi}, ϕ, D_p from IPI and Non-IPI Subjects
Using the Same Posttest of IPI Math,
D-Subtraction (N = 20 items)

	Posttest (IPI)			Posttest (Non-IPI)		
	1. r_{pbi}	2. ϕ	3. D_p	4. r_{pbi}	5. ϕ	6. D_p
1		.54	.53	.36	.43	.43
2			.71	.00	.10	-.07
3				.35	.41	.31
4					.95	.93
5						.91

Note: 5% = .444, 1% = .561

Table 5

Descriptive Statistics of Subjects
Taking Test Form A

Objective	No. of Items	Grade 3 N=49 Mean	St. Dev.	KR20	Grade 4 N=59 Mean	St. Dev.	KR20
I	4	3.47	.92	.62	3.90	.30	-.05
II	4	2.10	1.86	.95	2.78	1.57	.87
III	4	1.92	1.91	.97	2.66	1.70	.92
IV	4	1.33	1.68	.92	1.95	1.78	.91

Table 6

Reliability Coefficients Computed from KR20 and Their Corresponding
Criterion-Referenced Reliability Coefficients for Test Form B

Objective	No. of Items	Mastery Level	Grade 2 N=69 Mean	St. Dev.	KR20	r_c^2	Grade 3 N=110 Mean	St. Dev.	KR20	r_c^2
I	4	75%	2.49	1.49	.80	.82	3.46	.99	.71	.76
II	4	75%	.28	.82	.83	.98	2.75	1.52	.84	.85
III	4	75%	.14	.46	.47	.99	2.53	1.40	.72	.74
IV	4	75%	.07	.31	.39	.99	2.15	1.48	.74	.80

Table 7

Intercorrelations of r_{pbi}, ϕ, and D_p from
Grade 3 and Grade 4 Subjects
Using Test Form A
(N = 16 items)

	Grade 3			Grade 4		
	1. r_{pbi}	2. ϕ	3. D_p	4. r_{pbi}	5. ϕ	6. D_p
1		.91	.92	.70	.88	.44
2			.94	.49	.72	.25
3				.69	.73	.53
4					.80	.92
5						.52

Note: 5% = .497, 1% = .623

Table 8

Intercorrelations of r_{pbi}, ϕ, and D_p from
Grade 2 and Grade 3 Subjects
Using Test Form B
(N = 16 items)

	Grade 2			Grade 3		
	1. r_{pbi}	2. ϕ	3. D_p	4. r_{pbi}	5. ϕ	6. D_p
1		.52	.50	.28	.26	.19
2			.96	.28	.24	.20
3				.40	.29	.24
4					.92	.87
5						.97

Note: 5% = .497, 1% = .623

Table 9

The Intercorrelations of r_{pbi}, ϕ, and D_p
When Items Vary in Difficulty

	$r_{pbi} \times \phi$			$\phi \times D_p$			$r_{pbi} \times D_p$		
	High	Middle	Low	High	Middle	Low	High	Middle	Low
Sample 1	.92*	.77*	.12	.83*	.99*	.79*	.76+	.79*	.38
Sample 2	.73*	.68*	-.42	-.45	.81*	.82+	-.32	.88*	.14
Sample 3	.95*	.91*	.80*	.69+	.98*	.92*	.70+	.91*	.93*
Sample 4	.46	.93*	.45	.70*	.99*	.98*	.50+	.88*	.44

+ - Significant at 5%

* - Significant at 1%