R E P O R T R E S U M E S

ED 010 517 24

A COMPARISON OF ITEM SELECTION TECHNIQUES FOR NORM-REFERENCED AND CRITERION-REFERENCED TESTS.

BY- COX, RICHARD C. VARGAS, JULIE S.
PITTSBURGH UNIV., PA., LEARNING RES. AND DEV. CTR.

REPORT NUMBER BR-5-0253-REPRINT-7 PUB DATE FEB 66
EDRS PRICE MF-$0.09 HC-$0.64 16P.

DESCRIPTORS- *ITEM ANALYSIS, *TEST CONSTRUCTION, TEST VALIDITY, TESTING, *TEST SELECTION, POST TESTING, PRETESTING, DIAGNOSTIC TESTS (EDUCATION), MEASUREMENT INSTRUMENTS, *INDIVIDUAL INSTRUCTION, *DISCRIMINANT ANALYSIS, RESEARCH AND DEVELOPMENT CENTERS, PITTSBURGH, PENNSYLVANIA

AN INVESTIGATION WAS MADE TO DETERMINE TO WHAT EXTENT TWO METHODS OF ITEM ANALYSIS - NORM REFERENCED AND CRITERION REFERENCED - YIELD THE SAME RELATIVE EVALUATION OF TEST ITEMS. ITEMS WHICH DISCRIMINATED WELL BETWEEN STUDENTS SCORING HIGH AND LOW ON POST-TESTS WERE STUDIED TO SEE IF THEY ALSO DISCRIMINATED WELL BETWEEN PRETRAINING AND POST-TRAINING GROUPS. TWO SETS OF INDEXES WERE COMPUTED FOR THE ITEMS ON EACH OF TWO ARITHMETIC TESTS (ADDITION AND MULTIPLICATION), WHICH HAD BEEN GIVEN BOTH AS PRETESTS AND POST-TESTS IN AN INDIVIDUAL INSTRUCTION PROGRAM IN A PUBLIC ELEMENTARY SCHOOL. IT WAS FOUND THAT THE METHOD OF ITEM ANALYSIS ATTEMPTED IN THIS STUDY (PRETEST AND POST-TEST METHOD) SEEMS TO PRODUCE RESULTS SUFFICIENTLY DIFFERENT FROM TRADITIONAL METHODS TO WARRANT ITS CONSIDERATION WHEN CRITERION REFERENCED TESTS ARE DESIRED. TRADITIONAL ITEM ANALYSIS PROCEDURES WERE DEEMED APPROPRIATE IN THE SELECTION OF NORM REFERENCED MEASURES. (GD)

UNIVERSITY OF PITTSBURGH - LEARNING R & D CENTER

REPRINT 7

A COMPARISON OF ITEM SELECTION TECHNIQUES FOR NORM-REFERENCED AND CRITERION-REFERENCED TESTS

RICHARD C. COX AND JULIE S. VARGAS

A COMPARISON OF ITEM SELECTION TECHNIQUES FOR NORM-REFERENCED AND CRITERION-REFERENCED TESTS

Richard C. Cox

Julie S. Vargas

Learning Research and Development Center

University of Pittsburgh

February, 1966

A COMPARISON OF ITEM SELECTION TECHNIQUES FOR NORM-REFERENCED AND CRITERION-REFERENCED TESTS[1]

Richard C. Cox

Julie S. Vargas

University of Pittsburgh

The distinction between norm-referenced and criterion-referenced tests has been noted in several theoretical discussions of achievement measurement (Coulson and Cogswell, 1965; Ebel, 1965; Glaser, 1963). A norm-referenced measure indicates the relative standing of an individual in some norm group. Percentiles or grade equivalents, for example, are norm-referenced scores which compare an individual with national or local norms. Norm-referenced tests do not provide much information concerning the amount or kind of material a student has mastered.

Criterion-referenced tests, on the other hand, provide information in terms of specific behaviors mastered, without reference to the performance of other pupils. A score of 80 per cent, for example, indicates that an individual has successfully mastered 80 per cent of the behaviors specified on the test.

---

[1] Paper read at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois, February 1966.

The research and development reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education and Welfare under the provisions of the Cooperative Research Program.

The relative standing provided by norm-referenced tests is the information necessary for selection or grading purposes. Differences between individuals are maximized to better differentiate among those taking the test. Item selection procedures currently employed fulfill precisely this function; they select items which maximize differences among individuals.

In criterion-referenced tests, unlike norm-referenced tests, the purpose is not to discriminate between individuals but to discriminate between successive performances of one individual.

Criterion-referenced tests are needed in programs of individualized instruction in which students progress at their own rates along a sequenced program covering specific skills. The information required for assessment of pupil progress is curriculum-specific and does not concern the achievement of other pupils. When work is grouped into units, tests are commonly given to students before beginning the unit as diagnostic measures to identify those skills which are already mastered and those needing work. Occasionally a student does well enough on a pretest to skip a unit altogether. Criterion-referenced tests, therefore, should indicate whether or not a student will benefit by training in a unit, and should provide the basis for diagnosing the student's strengths and weaknesses.

The usual item selection techniques do not produce tests which indicate the value of a course of study for each student. These methods of item analysis tend to produce homogeneous tests, culling out items which are not similar to the majority of items. The

discarded items, however, may be covering important objectives or may be especially valuable for purposes of diagnosis. The problem is to find procedures for item analysis which will identify items which discriminate between those needing training and those not needing training on the skill covered by each item.

The usual item analysis procedures discriminate well--but for criterion-referenced tests, between the wrong groups. Instead of using high and low groups on total score for analysis, pre and posttest groups could be used for computation of difference indices. This would identify items which best discriminate between pre and posttest groups, indicating items which are most useful for pretest diagnosis.

To take an extreme example, any item which discriminates perfectly between pre and post-training groups must, by usual item analysis procedures, be rated as completely non-discriminating. For an item to discriminate perfectly by the first measure, all students must fail the item on the pretest but pass it on the posttest. An item which everyone passes after training, however, is answered alike by high and low scorers and, therefore, does not differentiate between the two groups commonly used for item analysis.

Thus items which discriminate perfectly, or nearly so, among pre and posttest groups necessarily will discriminate poorly among the post-trainees. This is a specific example in which the

two methods of item analysis give opposite assessments of an item.[1] In most cases, however, the two methods could be expected to rate items similarly, since generally difficult items are more likely to be passed by posttraining groups and by high posttest scorers than by pretrainees and by low posttest scorers.

The question asked in the present study is, then, to what extent the two methods of item analysis yield the same relative evaluation of items. Generally speaking, do items which discriminate well between students scoring high and low on posttests also discriminate well between pre and post-training groups?

Procedure

Two sets of indices[2] were computed for the items on each of two arithmetic tests, which had been given both as pretests and posttests in an individual instruction program in a public elementary school.

Because of the nature of the individualized instruction program, children from several grades took each test. Fifty children from grades one through four took the 31-item test in addition and 25 children from grades four through six took the 40-item multiplication test.

---

1 The reverse is not true, however. Maximum discrimination among post-trainees does not necessitate zero or low discrimination between pre and posttest groups. Similarly a zero discrimination index by either method has no implications for the discrimination value of the item by the other method.

2 For computation of the two indices, see the footnotes on Table 1.

Results

The rank ordering of items by discriminating power according to the two indices is presented in Tables 1 and 2. While for many items in both tests the indices rank the item similarly, a few noteworthy discrepancies exist. Item 24, from the addition test, for example, ranked first in discriminating among pre and posttest groups, but discriminated poorly between high and low posttest scorers.

The two ranks for each item are listed in Tables 3 and 4. The overall trend is toward agreement in the relative assessment of items by cheir discriminating power, as is indicated by positive, but small, rank order correlations for both tests.[1]

Table 5 shows the percentages of common items in the final tests constructed by taking the best discriminating items according to the two indices. If the final addition or multiplication test is to consist of the best two-thirds of the items from the item pool, approximately 3/4 to 4/5 of the items will be the same no matter which item discrimination index is used. It should be noted, however, that items which do not contribute substantially to discriminations between pre and posttest groups would be retained by usual item assessments. Similarly, some of the best discriminations between pre and posttest groups are made by items which will be eliminated by using the conventional upper-lower 27% method. The

---

[1] Both ρ correlation coefficients are statistically significant at the .01 level.

addition item previously mentioned, for example, which best discriminates between pre and post-training groups would be discarded following conventional item analysis procedures.

Implications

When deciding upon item analysis procedures the purpose of a test should be considered. If selection or grading of pupils is desired, a norm-referenced measure is required and tradition item analysis procedures are appropriate. Where criterion-referenced tests are desired, however, an alternate method of item selection is suggested. This new method evaluates items according to their ability to determine whether or not training would profit the student. The traditional and new pretest-posttest methods in the present study have been found to produce tests containing many of the same items. When about 1/3 of the items in the item pool were discarded, some items which are highly desirable for criterion-referenced tests were discarded by the traditional methods. This result, while logically expected for items with extremely high pretest-posttest discriminating power (difference indices of 90 or higher), occurred even though no such extremes were obtained. The highest pretest-posttest difference indices obtained on the two tests were 60 and 64, clearly not extreme values. In the practical situation, then, the method of item analysis suggested here seems to produce results sufficiently different from traditional methods to warrant its consideration when criterion-referenced tests are desired.

References

Coulson, J. E. and Cogswell, J. F. Effects of individualized instruction on testing. Journal of Educational Measurement, 1965, 2, No. 1.

Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1965, 22, 15-25.

Glaser, R. Instructional technology and the measurement of learning outcomes: some questions. American Psychologist, 1963, 18, No. 8.

Table 1

Rank Order of Addition Items By 1) The Difference Index, D, and 2) The Pretest-Posttest Difference Index, D_{pp}

Total number of students = 50

Item Number	D^1	Item Number	D_{pp}^2
13	69	24	60
27	61	20	52
26	54	25	52
14	46	26	52
23	46	27	52
28	46	19	50
2	38	28	50
18	38	29	50
29	38	18	46
30	38	30	44
11	31	21	42
15	31	17	36
16	31	23	30
20	31	12	26
22	31	16	20
25	31	31	20
31	30	3	18
12	23	11	18
17	23	4	16
19	23	8	16
21	23	14	12
4	15	1	10
8	15	5	10
10	15	10	10
24	15	13	10
5	8	6	8
1	0	7	6
6	0	9	6
7	0	15	2
3	- 8	2	0
9	- 8	22	- 4

[1] D = Difference Index; The percentage of students in the highest 27% in total posttest score who pass the item minus the percentage in the lowest 27% who pass the item.

[2] D_{pp} = Pretest-Posttest Difference Index; The percentage of students who pass the item on the posttest minus the percentage who pass the item on the pretest.

Table 2

Rank Order of Multiplication Items By 1) The Difference Index, D and 2) The Pretest-Posttest Difference Index, D_{pp}

Total number of students = 25

Item Number	D^1	Item Number	D_{pp}^2
19	62	16	64
28	62	11	56
16	50	14	52
27	38	12	48
30	38	15	48
38	38	10	44
4	37	17	44
5	37	28	44
6	37	13	40
14	37	18	40
15	37	20	36
17	37	25	36
18	37	26	36
20	37	22	32
22	37	23	32
24	37	27	32
40	37	19	28
3	25	33	24
12	25	38	24
13	25	24	20
23	25	39	20
25	25	8	16
29	25	29	16
34	25	35	16
26	13	7	12
32	13	30	12
11	12	34	12
39	12	37	12
1	0	4	8
2	0	5	8
8	0	9	8
9	0	21	8
21	0	40	8
31	0	1	4
33	0	2	4
35	0	3	4
37	0	6	4
7	-12	31	- 4
10	-12	36	- 4
36	-13	32	- 8

[1] D = Difference Index; The percentage of students in the highest 33% in total posttest score who pass the item minus the percentage in the lowest 33% who pass the item.

[2] D_{pp} = Pretest-Posttest Difference Index; The percentage of students who pass the item on the posttest minus the percentage who pass the item on the pretest.

Table 3

Rank of Each Item on Two Item Discrimination Indices
Addition Test

Item Number	Rank on D	Rank on D_{pp}
1	28.0	23.5
2	8.5	30.0
3	30.5	17.5
4	23.5	19.5
5	26.0	23.5
6	28.0	26.0
7	28.0	27.5
8	23.5	19.5
9	30.5	27.5
10	23.5	23.5
11	13.5	17.5
12	19.5	14.0
13	1.0	23.5
14	5.0	21.0
15	13.5	29.0
16	13.5	15.5
17	19.5	12.0
18	8.5	9.0
19	19.5	7.0
20	13.5	3.5
21	19.5	11.0
22	13.5	31.0
23	5.0	13.0
24	23.5	1.0
25	13.5	3.5
26	3.0	3.5
27	2.0	3.5
28	5.0	7.0
29	8.5	7.0
30	8.5	10.0
31	17.0	15.5

The Spearman rank order correlation between rank on the Difference Index and rank on the Pretest-Posttest Difference Index is .37 (significant at the .01 level).

Table 4

Rank of Each Item on Two Item Discrimination Indices
Multiplication Test

Item Number	Rank on D	Rank on D_{pp}
1	33.0	35.5
2	33.0	35.5
3	21.0	35.5
4	12.0	31.0
5	12.0	19.0
6	12.0	35.5
7	38.5	26.5
8	33.0	23.0
9	33.0	31.0
10	38.5	7.0
11	27.5	2.0
12	21.0	4.5
13	21.0	9.5
14	12.0	3.0
15	12.0	4.5
16	3.0	1.0
17	12.0	7.0
18	12.0	9.5
19	1.5	17.0
20	12.0	12.0
21	33.0	31.0
22	12.0	15.0
23	21.0	15.0
24	12.0	20.5
25	21.0	12.0
26	25.5	12.0
27	5.0	15.0
28	1.5	7.0
29	21.0	23.0
30	5.0	26.5
31	33.0	38.5
32	25.5	40.0
33	33.0	18.5
34	21.0	26.5
35	33.0	23.0
36	40.0	38.5
37	33.0	26.5
38	5.0	18.5
39	27.5	20.5
40	12.0	31.0

The Spearman rank order correlation between rank on the Difference Index and rank on the Pretest-Posttest Difference Index is .40 (significant at the .01 level).

Table 5

Percentage of Most Discriminating Items Selected In Common by Two Item Discrimination Indices for Different Lengths of Final Tests

Final Test Length as Proportion of Item Pool	Addition	Multiplication
one third	60%	23-53%*
two thirds	81%	74-78%*
Items discarded in common in bottom one third	60%	46-54%*

* The two values are the minimum and maximum overlap which could be obtained for the final form of the test by selecting among items which had tied ranks.