DOCUMENT RESUME

ED 074 147 TM 002 512

AUTHCR	Roudabush, Glenn E.
TITLE	Item Selection for Criterion-Referenced Tests.
PUB CATE	Feb 73
NOTE	16p.; Paper presented at annual meeting of the American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Criterion Referenced Tests; Diagnostic Tests; Educational Objectives; *Item Analysis; Post Testing; Pretesting; Speeches; Tables (Data); Technical Reports; *Test Construction; Test Validity
IDENTIFIERS	California Achievement Tests; Prescriptive Reading Inventory

ABJTRACT

The desirable characteristics of criterion referenced test items and sets of items are described. A two-stage item tryout and item selection procedure are also described. The paper presents the results of using the procedure as compared with traditional item selection procedures used in selecting items for norm.referenced tests. It was found that the items selected from the same item pool by the two procedures differ markedly. A rationale for these differences is presented and recommendations for appropriate uses of the two kinds of instruments are given. (Author)

ITEM SELECTION FOR CRITERION-REFERENCED TESTS

by

Glenn E. Roudabush

CTB/McGraw-Hill

A paper presented at the

American Educational Research Association meetings in New Orleans

February, 1973

ITEM SELECTION FOR CRITERION-REFERENCED TESTS

A criterion-referenced test is constructed to provide information on the performance of an examinee on a set of coherent objectives, usually in terms of mastery or non-mastery of each objective represented in the test. The objectives represented in the test will be directly or indirectly related to some curriculum or segment of curriculum. In mathematics, for example, the objectives may represent what is generally taught in fourth grade general math or specifically what is taught in a particular fourth grade math program. Or the objectives may represent what is to be taught in a six week course in life saving and water safety. Even if the objectives are not directly representative of a defined curriculum as, for example, in National Assessment, there is an implication that the objectives represent some behavior that the examinee is expected to have learned - usually in a formal school situation. A criterion-referenced test, then, begins with a set of objectives representing some curriculum and ends with reporting performance on each of those objectives. The characteristics of criterion-referenced tests derive from this curriculum orientation.

A criterion-referenced test is intended to supply information about the standing of an examinee with respect to a defined or implied curriculum. If the test represents a reasonably long span of the curriculum, it will yield many scores - one for each objective covered by the test. There is not much interest or value in the total test score, since it tells you little about the specific achievements or deficiencies of the examinee. This is an obvious and major difference between criterion-referenced tests and norm-referenced tests. A norm-referenced test provides information about the standing of the examinee with respect to a reference or norm group and this can be accomplished with a single aggregate total score. The total score in itself has little meaning except as a gross measure of amount of achievement in a given area. Meaning for the score is derived from the norm group, just as the criterion-referenced scores derive their meaning from the curriculum represented. A good criterion-referenced test should discriminate well between mastery and nonmastery of the objectives making up the curriculum of interest, just as a good norm-referenced test should discriminate well between examinees who have differing amounts of achievement in the general area of interest. This has implications for the way in which items are prepared and selected. Items in a criterion-referenced test should be sensitive to instruction; items in a norm-referenced test should be sensitive to individual differences.

A criterion-referenced test is generally intended to be diagnostic and prescriptive. The test should (1) accurately reflect the examinee's standing with respect to the curriculum, that is, show his specific strengths and weaknesses, (2) accurately reflect changes when the examinee's capability to perform has changed, and (3) lead to appropriate decisions for the further instruction of the examinee. A norm-referenced test, on the other hand, is generally intended to be descriptive and predictive. It should (1) accurately reflect the examinees standing with respect to the norm group, that is, show his relative position on the underlying quantity or trait being measured, and (2) accurately predict what the examinee will be able to do successfully. These distinctions lead to somewhat different views of reliability and validity for the two kinds of instruments. The usual validity and reliability coefficients reported for standardized norm-referenced tests have marginal utility for describing criterion-referenced tests. A criterion-referenced test should have demonstrable content validity and it should be sensitive to appropriate instruction. Reliability in the usual sense has less importance than the appropriateness of the decisions made that affect the treatment of the examinee.

This goes beyond the instrument itself and leads to considerations of minimizing risk or cost to the examinee.

Traditionally, for norm-referenced tests, test construction begins with some sort of comprehensive rationale describing the achievement domain or underlying trait intended to be measured and describing the kinds of items that should be written, frequently with examples. After the items are written, they are tried out on a sample of the target population. Item statistics are then computed including difficulty levels, point biserial correlations between each item and the remaining items, and some index of internal consistency, usually a KR-20. Items are selected that have difficulties around .5, so they will discriminate well between examinees, and that have high point biserials, so they will contribute to the homogeneity of the score. An attempt is also usually made to have the distribution of scores approximate a normal distribution. Normally distributed scores have valuable psychometric properties: they correlate well with other similar scores, provide meaningful derived scores, and so on. For a criterion-referenced test, these statistics are still important, but of less importance than the ability of the items to indicate mastery or nonmastery of particular objectives after instruction.

A criterion-referenced test begins with a set of coherent, clearly stated objectives. Each objective specifically describes the behavior that an examinee will be able to perform if he has mastered the objective, that is, each objective specifies a limited domain of behaviors. Items are then written for each objective that sample as purely as possible the specified domain of behaviors. This sample of behaviors will, of course, not be random, but hopefully, it will be representative of the domain. The items will then be tried out on a sample of the target population. Traditional item statistics will be computed and attention paid to them. It is more important, however, to determine if the items are sensitive to instruction. In order to do this, a two-stage item tryout is required, that is, a pre-instruction administration of the items followed by a period of time for instruction to occur, then a post-instruction administration of the items to the same students. It is also necessary to collect information as accurately as possible about the specific objectives appearing in the test that were taught to between the pre-instruction administration and the post-instruction administration of the items. If the instructional program is under the control of the test constructor, this information is relatively easy to obtain. If not, it can be approximated by asking the teachers what they have taught.

In order to select items that are sensitive to instruction, it is valuable to have some procedure for organizing the data and some numerical index reflecting each item's sensitivity. At CTB/McGraw-Hill, we have adopted a procedure described by Marks and Noll (1967) developed for a somewhat different purpose. First we obtain a two-by-two table of frequencies for each item at pre- and post-test like this:

Post-test

Pre-test	0	1	
0	f_1	f_2	$f_1 + f_1$
1	f_3	f_4	$f_3 + f_4$
	$f_1 + f_3$	$f_2 + f_4$	N

Here the rows represent, respectively, failed and passed the item at pre-test and the colums represent failed and passed the item at post-test, so that:

f_1 = the frequency of cases that failed the item at both pre- and post-test,

f_2 = the frequency of cases that failed the item at pre-test, but passed it at post-test,

f_3 = the frequency of cases that passed the item at pre-test, but failed the item at post-test, and

f_4 = the frequency of cases that passed the item at both pre- and post-test.

$N = f_1 + f_2 + f_3 + f_4$ = the total number of cases that were administered the item at both pre- and post-test.

Marks and Noll assume that there is some fixed non-zero probability, p, that a student who does not know the answer to the item will guess the correct answer. The value of p is determined by the item only and does not vary from student to student nor from occasion to occasion for the same student, that is, they admit of no partial knowledge and assume that an examinee's responses are independent at pre- and post-test when he does not know the correct answer and fails to learn it. They also assume that the only possible result of exposure to instruction between pre- and post-test is that a student learn the correct answer to an item. They admit of no forgetting so that a non-zero frequency of f_3 is solely due to guessing. The "true" value of f_3 is zero. With these assumptions, they then reason that f_1, those people who failed the item at both pre- and post-test, is composed only of people who in fact do not know the answer after instruction. Therefore f_1 is equal to the probability of guessing wrong twice times the number of people in the sample who do not learn the answer, that is:

$$f_1 = (1-p)^2 \hat{f}_1 \quad , \qquad (1)$$

where $\hat{f}_1$ is the "true" number of people who do not learn. Similarly f_2, those people who failed the item at pre-test and passed it at post-test, is composed of the number of people who learned the correct response and guessed wrong at pre-test plus the number of people who did not learn but guessed right at the post-test and wrong at the pre-test, so that:

$$f_2 = (1-p)\ \hat{f}_2 + p(1-p)\ \hat{f}_1 \quad , \qquad (2)$$

where $\hat{f}_2$ is the "true" number of people in the sample who did not know at pre-test, but have learned by the post-test.

Next f_3, those people who passed the item at the pre-test but failed it at the post-test, is again composed solely of those who do not know nor learn the correct answer but who guessed correctly at the pre-test, that is:

$$f_3 = p(1-p)\ \hat{f}_1 \quad . \qquad (3)$$

Finally, f_4, those people who passed the item at both pre- and post-test, is composed (1) of all of the people who in fact know the correct response at both pre- and post-test, (2) the number of people who learned the answer and also guessed correctly at the pre-test, and (3) the number of people who did not know nor learn the answer, but who guessed correctly at both pre- and post-test, that is:

$$f_4 = \hat{f}_4 + p\ \hat{f}_2 + p^2\ \hat{f}_1 \quad , \qquad (4)$$

where $\hat{f}_4$ is the "true" number of people in the sample who know the correct answer at both pre- and post-test.

From equations (1) and (3):

$$p = \frac{f_3}{f_1 + f_3} \qquad (5)$$

and equations (1) through (4) form a consistent system so that solutions for the $\hat{f}_i$ can be found:

$$\begin{aligned} \hat{f}_1 &= \frac{(f_3 + f_1)^2}{f_1} \quad , \\ \hat{f}_2 &= \frac{(f_2 - f_3)\ (f_2 + f_1)}{f_1} \quad , \\ \hat{f}_3 &= 0 \quad , \\ \hat{f}_4 &= f_4 - \frac{f_3\ f_2}{f_1} \quad . \end{aligned} \tag{6}$$

A ratio:

$$s = \frac{\hat{f}_2}{\hat{f}_1 + \hat{f}_2} \tag{7}$$

can serve as an index of the degree to which examinees are selecting the correct response to the item as a function of the instruction received between pre- and post-test, that is, a sensitivity index. This index is simply the proportion of cases that missed the item on the pre-test and then got it right on the post-test after a correction for guessing has been applied.

This procedure was applied to data obtained in a two-stage item tryout for the Prescriptive Reading Inventory (PRI), a criterion-referenced reading test published by CTB/McGraw-Hill in the fall of 1972. Items were selected to measure 90 separate reading objectives and these were arranged in four overlapping levels of the test nominally spanning grades 1.5 through grade 6. Information about what had been taught to the students in the tryout sample was obtained from a questionnaire that was filled out by the teachers of these students at about the time of the post-instruction administration of the items. The questionnaire listed each objective represented in the test, written out in full, with spaces by them to mark one of "taught before the pre-test," "taught between the pre-test and the post-test," and "not yet taught." In many cases, the teachers marked both the "taught before" and the "taught between" categories for particular objectives giving rise to an additional "review" category. The item tryout data was divided into these four categories.

For each item, then, for each of these four categories and for each grade group to whom the item was administered (two or three grades), we computed the two-by-two table of frequencies, the corresponding table of proportions, the two-by-two table of corrected or estimated "true" frequencies, the corresponding table of proportions, and the sensitivity index. Since more than 1,600 items were tried out, this produced an enormous amount of data.

Theoretically, the value of the sensitivity index should be low for the "taught before the pre-test" group, higher for the "review" group, highest for the "taught between the pre-test and the post-test" group, and close to zero for the "not yet taught" group. In our case, we rarely had enough cases in more than one or two of the groups to get a stable value for the index. We feel that, in order to get a reasonably reliable value for the index, that there should be at least fifty cases who missed the item at the pre-test, that is, the sum of f_1 and f_2 should be fifty or more. The cases in the f_4 cell, those who passed the item at both the pre- and post-test, do not contribute to the calculation of the index and if the proportion of cases in the f_4 cell is high, which it generally is especially for the "taught before" and "review" groups, then the index will be of little value. Where we were able to partially validate the pattern of index values from group to group, it generally held up, except that the values for the "taught before" and "not yet taught" groups tended to be higher than expected and the values for the "review" and "taught between" groups tended to be lower than expected. This may, in part, be due to the unreliability of the questionnaire data, upon which the categorization depended.

Table 1 shows the results for the "taught between pre- and post-test" group for seven items, all of which were written to measure the objective "The student will be .'le to identify compound words." The first thing you will notice is that as many labels as possible were omitted to save space. Each 3 by 3 set of numbers is a two-by-two table with marginals organized as described above. The first one at the top of the page labelled "IF" is the observed frequencies. The second one down labelled "IP" is the proportions corresponding to the observed frequencies. The third labelled "IF (EST)" is the corrected or estimated "true" frequencies. The last labelled "IP (EST)" is the proportions corresponding to the estimated frequencies. At the very bottom is the sensitivity index labelled "D". These data are somewhat better than typical for first graders. It is rare, in our data, to find that all items for an objective have acceptable values for the sensitivity index. Look now at the marginal proportions in the second table for those who passed the item at pre-test and those that passed the item at post-test. These are the item difficulties at pre- and post-test respectively. For item 11, the first item in the table, the pre-test item difficulty is .58 and the post-test difficulty is .80. This is an additional indication of the sensitivity of the item.

The sensitivity index for reading tends to be higher at the lower grade levels and higher fo: discrete skills like recognizing compound words while it tends to be lower at the upper grades and for comprehension type items. This is, of course, to be expected, since reading tends to converge to a more or less unitary skill as practice accumulates.

Table 2 shows rather typical results for seven items all written to measure the objective "The student will be able to identify the root word in words with added endings that involve spelling changes". Notice that item 96, the next to the last item in the table, has a negative sensitivity index. This occurs whenever f_3, the number of cases who passed the item at the pre-test and failed it at the post-test, is larger than f_2, the number of cases who failed the item at the pre-test and passed it at the post-test. A negative index indicates that there is a serious problem with the item. In this case, there is nothing obviously wrong with the item, but looking at the pattern of frequencies compared to the other items in the set, it seems plausible that the item was miskeyed.

The upper limit of the index is one and it generally should not go below zero, though it obviously can and does. We had a few objectives the items for which all had negative index values. In one case, for an objective having to do with alliteration, the item writer had been unable to write items that got at the intent of the objective. We subsequently decided that the objective could not be reasonably measured in a paper and pencil test and excluded it from the published test. In other cases, the objective was misplaced and the items were grossly inappropriate for the students who completed them.

After selecting items using the sensitivity index as the primary criterion for selection, I ran several traditional item analyses lumping all the items from a tryout booklet together to see what items would have been selected in the traditional way. One set of items was related to vocabulary objectives and two others were all comprehension type items. In each case less than half of the items selected for the criterion-referenced test were selected for a hypothetical norm-referenced test. For the vocabulary test, 23 items were selected and of those, 10 were also used in the PRI while 13 were not. For one of the comprehension tests, 37 items were selected, 16 of which were included in the PRI. For the other, 42 items were selected, 18 of which were included in the PRI. Further, the objectives were unevenly represented in the hypothetical norm-referenced tests. Some objectives were not represented at all while others had as many as 8 or 10 items selected. Using sensitivity to instruction as the major criterion for item selection leads to choosing a different set of items than would ordinarily be chosen.

We also had scores for the California Achievement Tests, 1970 Edition, Reading Vocabulary subtest for our tryout sample. Using the set of 10 vocabulary related objectives, I obtained the intercorrelations of these with the CAT Reading Vocabulary scores and then did a stepwise regression analysis to see how well the CAT could be predicted from the objective scores. Table 3 shows the intercorrelation matrix. Note that the highest correlation of any objective is .48 with the vocabulary test. Generally they run about .40. The intercorrelations of the objectives with each other average around .50. The multiple correlation with all ten objective scores in regression only reached .55. This tends to show rather clearly that the two kinds of tests are not much alike and that scores on one might easily change without a corresponding change in the other.

Reference

Marks, E. and Noll, G. A. Procedures and criteria for evaluating reading and listening comprehension tests. Educ. and Psychol. Meas., 1967, 27, 335-348.

```
                    PRI PRE-POST TEST ANALYSIS BY QUESTIONNAIRE INFORMATION               DATE=12/13/71
LEVEL A, BOOKLET 3, GRADE 1, OBJECTIVE IV-9          N= 147                           TEST DATE= 2/15/71- 4/12/71
                                                                                                 (1.6-1.8)
          ITEM  11             ITEM  12             ITEM  13             ITEM  14             ITEM  15             ITEM  16             ITEM  17

        10    39    49      39    39    78      16    46    62      16    38    54      29    42    71      40    40    80      14    35    49
IF      14    55    69       8    32    40      11    45    56      12    52    64      10    37    47      10    28    38      15    54    69
        24    94   118      47    71   118      27    91   118      28    90   118      39    79   118      50    68   118      29    89   118

      0.08  0.33  0.42    0.33  0.33  0.66    0.14  0.39  0.53    0.14  0.32  0.46    0.25  0.36  0.60    0.34  0.34  0.68    0.12  0.30  0.42
IP    0.12  0.47  0.58    0.07  0.27  0.34    0.09  0.38  0.47    0.10  0.44  0.54    0.08  0.31  0.40    0.08  0.24  0.32    0.13  0.46  0.58
      0.20  0.80  1.00    0.40  0.60  1.00    0.23  0.77  1.00    0.24  0.76  1.00    0.33  0.67  1.00    0.42  0.56  1.00    0.25  0.75  1.00

        58    60   118      57    37    94      46    59   105      49    46    95      52    43    95      63    38   100      60    41   101
  IF     0     0     0       0    24    24       0    13    13       0    24    24       0    23    23       0    18    18       0    17    17
(EST)   58    60   118      57    61   118      46    72   118      49    69   118      52    66   118      63    56   118      60    58   118

      0.49  0.51  1.00    0.48  0.32  0.80    0.39  0.50  0.89    0.42  0.39  0.80    0.44  0.36  0.81    0.53  0.32  0.85    0.51  0.35  0.86
  IP  0.00  0.00  0.00    0.00  0.20  0.20    0.00  0.11  0.11    0.00  0.20  0.20    0.00  0.19  0.19    0.00  0.15  0.15    0.00  0.14  0.14
(EST) 0.49  0.51  1.00    0.48  0.52  1.00    0.39  0.61  1.00    0.42  0.58  1.00    0.44  0.56  1.00    0.53  0.47  1.00    0.51  0.49  1.00
           D =0.51             D =0.40             D =0.56             D =0.48             D =0.45             D =0.38             D =0.41
```

Table 1. Pre-test posttest item statistics for seven items measuring the objective "The student will be able to identify compound words."

```
                         PPI PRE-POST TEST ANALYSIS BY QUESTIONNAIRE INFORMATION                DATE=12/22/71
LEVEL 3, BOOKLET 1, GRADE 2, OBJECTIVE IV-10        N= 242                                    TEST DATE= 2/ 8/71- 5/11/71
                                                                                                        (2.6-2.9)
            ITEM. 91              ITEM  92              ITEM  93              ITEM  94              ITEM  95              ITEM  96              ITEM  97

          44    49    93       27    34    61       22    40    62       61    49   110       80    48   128       27    34    61       58    61   119
 IF       21    32   103       24   111   135       24   110   134       31    55    86       25    43    68       57    78   135       22    55    77
          65   131   196       51   145   196       46   150   196       92   104   196      105    91   196       84   112   196       80   116   196

        0.22  0.25  0.47     0.14  0.17  0.31     0.11  0.20  0.32     0.31  0.25  0.56     0.41  0.24  0.65     0.14  0.17  0.31     0.30  0.31  0.61
 IP     0.11  0.42  0.53     0.12  0.57  0.69     0.12  0.56  0.68     0.16  0.28  0.44     0.13  0.22  0.35     0.29  0.40  0.69     0.11  0.28  0.39
        0.33  0.67  1.00     0.26  0.74  1.00     0.23  0.77  1.00     0.47  0.53  1.00     0.54  0.46  1.00     0.43  0.57  1.00     0.41  0.59  1.00

          96    41   137       96    19   115       96    33   130      139    27   166      138    30   166      261   -72   190      110    54   164
 IF        0    59    59        0    81    81        0    66    66        0    30    30        0    28    28        0     6     6        0    22    22
(EST)     96   100   196       96   100   196       96   100   196      139    57   196      138    58   196      261   -65   196      110    76   186

        0.49  0.21  0.70     0.49  0.10  0.59     0.49  0.17  0.66     0.71  0.14  0.85     0.70  0.15  0.86     1.33 -0.37  0.97     0.56  0.27  0.83
 IP     0.00  0.30  0.30     0.00  0.41  0.41     0.00  0.34  0.34     0.00  0.15  0.15     0.00  0.14  0.14     0.00  0.03  0.03     0.00  0.16  0.16
(EST)   0.49  0.51  1.00     0.49  0.51  1.00     0.49  0.51  1.00     0.71  0.29  1.00     0.70  0.30  1.00     1.33 -0.33  1.00     0.56  0.44  1.00
          D = 0.30             D = 0.16             D = 0.26             D = 0.16             D = 0.18             D =-0.38             D = 0.23
```

Table 2. Pre-test posttest item statistics for seven items measuring the objective "The student will be able to identify root words that have added endings involving spelling changes."

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Best Word for Sentence	1.00	.55	.56	.50	.50	.58	.57	.58	.44	.50	.48
2. Most Precise Word for Sentence	.55	1.00	.50	.55	.42	.64	.60	.54	.49	.42	.40
3. Phrases and Definitions	.56	.50	1.00	.45	.53	.49	.49	.44	.41	.40	.37
4. Words in Context	.50	.55	.45	1.00	.43	.63	.57	.50	.50	.46	.41
5. Multi-meaning Words	.50	.42	.53	.43	1.00	.50	.46	.43	.39	.42	.36
6. Words and Definitions	.58	.64	.49	.63	.50	1.00	.67	.58	.48	.51	.40
7. Multi-meaning Words and Definitions	.57	.60	.49	.57	.46	.67	1.00	.57	.44	.54	.41
8. Synonyms	.58	.54	.44	.50	.43	.53	.57	1.00	.37	.57	.44
9. Antonyms	.44	.49	.41	.50	.39	.48	.44	.37	1.00	.30	.27
10. Homonyms	.50	.42	.40	.46	.42	.51	.54	.57	.30	1.00	.39
11. CAT Reading Vocabulary	.48	.40	.37	.41	.36	.40	.41	.44	.27	.39	1.00

Table 3. Intercorrelations of ten vocabulary related objective scores and the California Achievement Tests Reading Vocabulary subtest.