DOCUMENT RESUME

ED 108 932 SE 019 164

AUTHOR Rueve, Mary H.; Grate, John H.
TITLE Revisions and Research Design for a Computer-Based Criterion-Referenced Testing Program, Model for Improving Basic Skills.
PUB DATE 75
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Basic Skills; *Criterion Referenced Tests; Educational Accountability; Elementary Education; Elementary School Mathematics; *Item Analysis; *Mathematics Education; Norm Referenced Tests; *Test Construction; Testing Programs
IDENTIFIERS Cincinnati Mathematics Inventories; Research Reports; *Test Revision

ABSTRACT

This paper describes the methods used for revising the Cincinnati Mathematics Inventories, a battery of criterion-referenced tests of basic skills used in the Cincinnati Title III program and other city-wide special programs. Each of these tests is designed to cover a half-year of work; items are included to sample skills listed in the Catalog of Basic Skills used in the system. The tests are used in grades one through eight. The procedure by which the tests were revised involved twelve sequential steps. Skills listed in the catalog were revised, and items were individually examined to test their coherence with the revised skills. Item statistics (r-biserial and a difficulty index) were computed and negative discriminators were eliminated. Teacher criticisms of items were used in revising or discarding items. Items were compared on the basis of statistical properties with others from the same skill areas or strands to determine acceptability. Four criteria were used in selecting among items remaining in the pool. The author describes this method as an unorthodox combination of criterion-referenced and norm-referenced techniques which is dynamic, responsive to change, and acceptable to educational decision makers in Cincinnati. (SD)

# Revisions and Research Design for a Computer-Based Criterion-Referenced Testing Program, Model for Improving Basic Skills

Mary H. Rueve
John H. Grate

Cincinnati Public Schools

Paper presented as part of a demonstration and discussion of the Model for Improving Basic Skills Project to the American Educational Research Association Convention, Spring 1975.

Now that you have criterion-referenced tests, what do you do with them? How do you revise? Methods for revising norm-referenced tests are well established; methods for revising and improving criterion-referenced tests are the subject of great debate. The problem is further complicated by the fact that what may be appropriate for one type of criterion-referenced test may not be appropriate for another. Despite these difficulties, large city school systems are adopting criterion-referenced tests as part of instructional management systems (Grosswald, 1973). The Cincinnati Public Schools is one such system.

This paper outlines the procedures used to revise our criterion-referenced tests, the Cincinnati Mathematics Inventories[1] now being used in ESEA Title III project, Model for Improving Basic Skills, and special Cincinnati Public School programs.

Trying to capture the best of both worlds, in our revision process we used a combination of criterion- and norm-referenced techniques. We continued the criterion-referenced philosophy followed in the development of the Inventories, that each item must measure a clearly defined skill or objective. We relied heavily on content validity (based on the judgment of our curriculum experts) to show that the skill content is valid and that the item measures the skill. We also used the traditional norm-referenced statistics to look further at items. Using a combination of techniques possibly sacrifices something in logical consistency but it increases the usability of the inventories in the instructional program.

[1] These tests were developed by the Planning and Development Branch of the Research and Development Department, Cincinnati Public Schools.

## The Criterion

The Catalog of Mathematics Skills is a listing of mathematics skills or objectives for grades one through eight. The Catalog is divided into concept areas or strands; each strand contains skills arranged in sequential order. Examples of how each skill is to be measured is given with the skill description.

The content and placement of skills are based on textbooks, CAI programs, and the Cincinnati curriculum. Appendix 1 is a page from the Catalog.

## The Tests

The Cincinnati Mathematics Inventories are a series of tests, each covering a half year or five skills from each concept area or strand covered at that grade level. Each item on an inventory is a sample of a skill in the Catalog. (Appendix 2 shows the relation of items to skills.) A more detailed description of the Cincinnati Mathematics Inventories is contained in the Teachers Handbook: Mathematics (1975) and the Cincinnati Instructional Management System (1975).

## Revision Procedures

STEP 1: Revise skills.

The skills in the Catalog were revised and modified before any revisions were made to the tests since any changes in a criteria should result in a change in the corresponding item.

STEP 2: Does item fit skill in Catalog?

Each item was checked against the skill it was measuring to determine whether the item was a valid sample of the skill. If the item was not judged as a valid measure of the skill, it was eliminated. If the item passed this first step, then the next step was to go to the data analysis.

Data Analysis. For each inventory, item responses were pooled at the end of the year. Two test statistics were computed, nonspurious point biserial correlation and the index of item difficulty. The nonspurious point biserial correlation coefficient shows the relation between an item and the total test score (item-to-total correlation) with that item not included in the total score. In other words, the correlation coefficient indicates to what extent success on the item is related to success on the test as a whole or the extent to which students who did well on the whole test did better on this particular item than students who did poorly on the whole test. Of course, the higher the correlation, the closer the relation. A negative correlation coefficient indicates that students who did well on the test as a whole tended to miss this item and students who did poorly on the whole test tended to get this item correct. In this case the item is called a "negative discriminator" and is considered a defective or undesirable item.

The other test statistic, the index of item difficulty (p value) tells the proportion of students who answered the item incorrectly. A difficulty level of .89 means that 89% of the students gave an incorrect response to that item. Appendix 3 shows a sample of the data analysis.

STEP 3: Is the item a negative discriminator?

Any item which was a negative discriminator was eliminated. This was the only instance where the data analysis was used without also considering the content of the item.

STEP 4: Was the item criticized by teachers?

Teachers who were using the tests in their classrooms checked for items which were confusing or inappropriate and wrote suggestions for changing the objectionable items. This was a particularly valuable step since teachers administering the inventories had an insight into what was confusing to children. We also found that on a poor item teachers often explained the item to the children and in this way contaminated the data analysis. Criticized items were either revised or eliminated.

In interpreting and using the data analysis from this point on, the assumptions peculiar to criterion-referenced tests were considered. The main concern was to guard against rejecting a good item which actually did measure a desirable and appropriate skill even though the test statistics might indicate that the item was not significantly correlated to total score or that the item was not discriminating. In other words the statistics were used to red flag items which might be poor; a value judgment rather than an arbitrary rule based on the data analysis was followed.

STEP 5: Is the correlation of item-to-total score significant?

If the correlation of the item-to-total score was significant, the item passed this step. If the correlation was nonsignificant, then the item was examined in two ways to determine the cause of the low correlation. First the item was checked to determine whether it was confusing, poorly written, or asking for obscure information, etc. If the item was defective it was of course rejected.

STEP 6: If item-to-total correlation is nonsignificat, is this typical of items on that strand?

If the item was not considered defective, then the item-to-total correlations of other items on the same strand were examined. If these correlation coefficients were also low or nonsignificant, then we went back to our mathematics supervisors. In consultations with the supervisors we found that there were certain concept areas (strands) which should be taught, but which our teachers were frequently skipping. If the item was from a strand of this nature, it was not rejected because of a nonsignificant correlation.

STEP 7: Is the index of difficulty extreme?

In using the index of difficulty for each item, again criterion-referenced assumptions were kept in mind. Items were not rejected only for being extremely easy or difficult. If an item was very easy it was checked to be certain that it was an appropriate sample of the skill it was measuring, that no clues were included, and that the distractors were plausible. Extremely low difficulty levels for items occurred most often at the primary levels indicating that teachers at this level were holding their students accountable for mastery of skills before they moved to the next inventory.

The extremely difficult item was treated in a similar manner to the item with a nonsignificant item-to-total correlation--it was checked to determine why it was difficult; if it was a word problem, was the vocabulary too advanced, etc.

STEP 8. If only one item measures a skill, retain that item.

If only one item measuring a skill passed the first six steps then that item was retained. The item was considered to have passed the first six steps even if it had been modified in the process and was now judged satisfactory.

Selection Among Items

In the revision of the Inventories it was decided to have a skill at each level of a strand and only one item to measure each skill. Steps 9 and 10 refer to the methods of selecting among items measuring the same skill. Of course only items which had passed the first six steps were considered.

STEP 9: Does the difficulty level of one item "fit" better than others?

If there were several acceptable items on a strand, then the item with a difficulty level which fit was selected. (Items on a strand are to become more difficult as the skill levels increase.)

STEP 10: Select item with highest item-to-total correlation.

If no item was selected at Step 9, then the item with the highest item-to-total correlation was selected as the item to measure that skill.

STEP 11: Are item difficulty levels out of order?

If the item difficulty levels on a strand were out of order, then the skills and items were rearranged. This was done of the basis of form 1. Then the other two forms were modified in the same way.

STEP 12: If no item measures a skill, write new item for it.

If a skill was added to the Catalog or if the item(s) measuring a skill was eliminated rather than revised, then an item was written to measure that skill.

Since in this item revision process, new items were added and old ones modified, it will be necessary to go through a similar revision process again.

## Summary

We are doing test development in an applied situation. A good amount of time and effort have gone toward involving supervisors, principals, and teachers in development of skill definitions and tests. Although some of our procedures have been unorthodox, less than "pure," we have a system which fits our needs, which is dynamic and responsive to change, and which decision-makers in Cincinnati have "bought into."

We have followed the policy that we will use any method of test construction or revision that increases the probability that our materials will be used as part of an instructional management system to help teachers teach and children learn.

## Bibliography

Grosswald, J. Testing perspectives in the large cities. NCME, 1973, Vol. 15, No. 3, 4.

Instructional Management System, Cincinnati Public Schools, Department of Research and Development, 1975.

Teachers Handbook: Mathematics, Cincinnati Public Schools, Department of Research and Development, 1975.