DOCUMENT RESUME

ED 193 272 TM 800 587

AUTHOR Noggle, Nelson L.
TITLE CRTs and NRTs Together.
PUB DATE Jun 79
NOTE 13p.: Paper presented at the Annual Conference on Large Scale Assessment (Denver, CO, June, 1979).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests: Computer Assisted Testing: *Criterion Referenced Tests: Diagnostic Tests: Elementary Secondary Education: *Norm Referenced Tests: Program Evaluation: *Testing Programs: Test Results
IDENTIFIERS Test Use: *Title I Evaluation and Reporting System

ABSTRACT

The potential use of criterion referenced tests (CRT) and norm referenced tests (NRT) in the same testing program is discussed. The advantages and disadvantages of each are listed, and the best solution, a merging, is proposed. To merge CRTs and NRTs into an overall testing program, meaningful and useful to all levels, consideration must be given to the role of both survey tests and diagnostic tests, including when and why they are administered. The possibility of having norms for a CRT is discussed. Methods of equating CRTs to NRTs are mentioned: for example, Model A from the Title I system recommends a shortened version of the equipercentile approach. A list of key recommendations for this procedure is offered. A futuristic look is taken at the proposed single evaluation process, its merging accomplished through both modern measurement and computer technology. (Author/GK)

CRTs and NRTs

Together

Nelson L. Noggle

Northwest Regional Educational Laboratory

Portland, Oregon

Paper presented at the Annual
Conference on Large Scale Assessment
Sponsored by the
National Assessment of Educational Progress
Denver
June, 1979

CRTs and NRTs: Together

Introduction

This paper was prompted by a need to continue discussion of the potential for criterion- and norm-referenced tests in the same testing program. Hambleton et al (1978), tended to clarify the CRT situation. The NRT has been clarified many times over. There still remains, however, a cloud for the everyday test user--which one, and why?

Some of the stimulation for writing this paper came from a personal desire to see efficient uses of both kinds of measurement. Another portion resulted from work being done with NRTs and CRTs for evaluating Title I programs. Since many of the Title I remedial programs are specified in terms of well-defined instructional objectives, CRTs are a natural selection for evaluation instruments. The problem, however, is that pending state and federal reporting requirements mandate the use of a test score which is based on NRTs. If a Title I project wants to use a CRT, they will be forced to also use an NRT.

The final reason for this paper was to acknowledge the information needs of different levels of educators, and to suggest that both CRTs and NRTs can be helpful to all levels. For example, a parent would find it helpful to know that the list of unlearned objectives for their child decreases as the school progresses. That same parent would like to know how their child compares over time to other students nationally. Likewise, the school board wants to know the same thing for the overall program. And, the U.S. Congress probably would like to know the same thing for federal programs.

Historical Perspective

Before and during the 1960s the typical achievement test was either a teacher-made test, an end-of-chapter test from a text, or a publisher's norm-referenced test. Each type of test was constructed, depending on available expertise and resources, in similar ways, except that the norm-referenced test (NRT) had an additional mission. The NRT was required to sort levels of achievement according to the relative achievement of those taking the test, i.e., it was required to provide a reference to the finely

tuned discrimination between students having very small differences in performance on the test. This resulted in test users comparing students or programs against various levels of performance from a reference group or norm, hence the term "norm-referenced".

In the early 60s, users of NRTs began calling for shorter tests. The amount of time spent for testing, as well as the cost of testing, was becoming too great in relation to the beneficial uses made of the tests. Many users filed away the test results and never used them at all. Some tried to use them to evaluate district, state and federal programs, while others attempted to use them to select students for programs. And, some tried to use them to counsel students and parents. Long tests, however, were not seen as being necessary to do these things.

In the mid-to-late 60s, users of NRTs began criticizing the tests for being biased against minority groups, or for being non-relevant to local instructional programs. Low scores were explained away as a lack of relevance.

At about this same time, educators began emphasizing the management of teaching via specified learning outcomes or objectives. Instructional materials and procedures abounded with lists of objectives. A parallel trend in testing began, which called for a greater focus on how much was learned rather than a comparison of different levels of learning. Soon, such terms as objective-referenced testing became household terms among educators, along with an increased desire to set mastery levels.

The unfortunate aspect of the trends in the late 60s and early 70s was the unnecessary pitting of CRTs vs. NRTs. Several states and districts replaced their NRTs with CRTs, while most established a "wait-and-see" stance about CRTs with lost confidence in their NRTs. Publishers, especially the larger ones, diverted development resources to react to the CRT market. The first emphasis was in the area of survey tests. Two kinds of published survey CRTs emerged, the shelf version and the custom version. The shelf version covered fewer objectives than their NRT counterparts, but used a few more items to measure each objective. The objectives in the shelf version were selected by the publisher in much the same way as they select them for their NRTs, with one major exception. Sometimes

objectives appeared in the CRTs that did not appear in the NRTs because of the requirement that NRT items must discriminate. However, most of the shelf CRTs still covered essentially the same content as the NRTs.

The custom or tailored version of the published CRTs typically contained a pool of objectives and items from which a potential user could select objectives, items, or both, and the publisher would print "locally relevant" CRTs. This process, while more expensive than shelf tests, was seen as being more relevant. We will not address in this paper the question if published CRTs can alwys be called CRTs, because they cannot.

There are two rather popular ideas of what comprises a CRT (Hambleton et al, 1978). The first requires that the test have a criterion for mastery. This is regarded by most users as a set percentage of items correct on a test (or on an objective) with the most common percentages being somewhere between 75 and 90. The second specifies that the test (or objective) should contain a representative sample of the tasks that make up the intended skill or knowledge to be tested. This latter idea is often associated with another testing term, domain-referencing. Typically, though, both ideas are considered when selecting or constructing a CRT.

Not all of the CRTs come from publishers. Many states and districts have constructed their own. Some local CRTs were built long before publishers were able to market them. Others were constructed as a way of guaranteeing local relevance, or to motivate local involvement and commitment to testing.

Another change has taken place which affects the future of CRTs and NRTs, alone or together. In 1974, Congress passed amendments to the Title I programs in the basic skills. As a result of that law, the U.S. Office of Education has developed an evaluation and reporting system, under contract with the RMC Research Corporation, Mountain View, California. The essence of this system is the reporting of NRT-based gain scores as determined by the difference between actual performance and an empirically based no-treatment expectation. Each of the evaluation designs suggested by this system allows for the use of CRTs, as long as the CRT results are converted to NRT-based scores for reporting purposes.

Merging CRTs and NRTs: Today

The historical perspective illustrated CRTs, as they are used primarily as survey tests. However, many of the CRTs being used are comprehensive diagnostic tests, or a series of diagnostic tests. When we look closely at testing practices today, especially at all levels of test usage, we must carefully delineate between those using or wanting survey CRTs and those using or wanting diagnostic CRTs. For example, administrators and curriculum specialists want "bottom-line" data. From the NRT they generally want to know basically three things:

1. How well did we do in comparison to others?
2. How well did we do in comparison to the past?
3. How does the picture look from subject to subject?

From the CRT they usually want to know the following:

1. How much do our students know?
2. Do they know as much as we expected?
3. What are the major strengths and weaknesses?
4. How well did we do in comparison to the past?
5. How does the picture look from subject to subject?

Several districts and a few states now use an NRT survey and a CRT survey to look at programs--to help answer questions like these.

Teachers, however, tend to want more information than a survey test can offer. The typical survey test does not provide them with what they really want, which seems to be:

1. Are the objectives the same ones that I teach?
2. How did each student do on each objective?
3. Which students need help on the same objective?
4. How can I verify if the remediation has worked?

CRTs used primarily as a posttest provide information which is too late to do teachers much good. CRTs used as a pretest can help, provided the scoring does not take too long. However, the CRTs used as a pretest are usually the survey variety and only cover the more important objectives--what about the many specific objectives needed to be learned on the way to the important ones?

The idea of using NRTs and CRTs together, however, is more than simply administering both. The programs that derive the most benefit seem to plan a total test program around either the questions that need answering or the decisions that need to be made. There seem to be two categories for those questions and decisions:

1. How well are we doing on what we say we are doing?
2. Are we doing and learning the right things? Are we in step?

To merge NRTs and CRTs, then, into an overall testing program which is meaningful and useful to all levels, careful consideration must be given to the role of both survey tests and diagnostic tests, not just NRTs and CRTs, including when and why they are administered. As various districts attempt to merge them, some commonalities become evident:

1. The survey test is an overall umbrella representing the major questions to be considered.
2. The NRT, for example, represents a benchmark for the district to compare itself in the major subject areas--are we in step?
3. The NRT can also help to determine how well the district is doing on a few of the more global objectives, depending on their correspondence to instruction and the way in which the NRT was constructed.
4. The survey CRT, however, represents the more definite check on learning of key local instructional objectives.
5. The diagnostic test takes up where the NRT and survey CRT leave off:
   a. it provides back-up information student by student and objective by objective in relationship to concerns raised by the survey tests.
   b. it covers objectives not covered by the survey tests.
   c. it is a more flexible tool for use day to day.

Let's take another approach. Let's look at a typical kind of reported data from a CRT--the percentage of students mastering an objective. Suppose we discovered from a survey CRT in Reading that 73

percent of the students completing fourth grade had mastered "main ideas". Assuming that the CRT was adequately constructed for such mastery determinations, the user knows only one thing--73 percent mastered it. To know more, the user needs a frame of reference for interpretation. One useful type of reference is a "desired", or expected, percentage of students as a criterion, which usually is determined subjectively rather than empirically. "Our goal is to bring 80 percent of the fourth-grade students to mastery of main ideas by the end of fourth grade." If 80 percent were our goal, then we would have to interpret 73 percent as falling short.

Another type of reference would be past data, of which there are two types: (1) past performance of the same students; and (2) past performance of similar students at the same grade level. Suppose we had both kinds of past data, and found that 48 percent of the same students were mastering main ideas at the beginning of the year, and that 64 percent of last year's fourth graders had mastered main ideas by the end of the year. We can now add more to our interpretation. While we were not able to meet our 80 percent goal, we were able to increase the percentage of fourth graders mastering main ideas from 48 percent to 73 percent, and reach a higher percentage, 73 percent over 64 percent, than we did with last year's fourth graders.

Another type of reference would be a comparison to larger reference groups, of which our fourth graders are a part. For example, suppose we are a school, and we found that 60 percent of the district's fourth graders mastered main ideas. Our interpretation would now include that we were better than the district as a whole, 73 percent over 60 percent. Then suppose we found that 85 percent of the nation's fourth graders had mastered main ideas by the end of fourth grade. Our interpretation, which was beginning to sound fairly successful, would suddenly change--we had not even reached the national average. Or, suppose we had found that only 50 percent of the nation's fourth graders had mastered main ideas--our success story continues to grow.

This sequence of CRT interpretations was done to illustrate an undeniable fact to be found in our schools, districts and states. Just because we use CRTs does not mean we are willing to give up referencing to past or present performance from various groups. The idea of comparing ourselves to the norm is still with us; but the idea of using something other than "average" performance as the only reference has been brought about by CRTs. While we are still interested in comparing to the district, state or national average, we have become more interested in higher criteria, goals, or standards; and we have become acutely aware of the building blocks that led us to those standards.

Adding Norms to CRTs

The possibility of having norms for a CRT is not unrealistic. At least one publisher advertises a normed CRT; and the Title I Evaluation and Reporting System requires that if you use a CRT, you must report in terms of NRT norms that are based on the relationship between the CRT and an NRT.

Several Methods of equating CRTs to NRTs are possible, each having its own merit in terms of utility, burden, and cost. Model A from the Title I system recommends a shortened version of the equipercentile approach, where the median CRT posttest score is given an NRT score based on an equating of the CRT and NRT at pretest time. Other approaches, such as those described by Angoff (1971), or the latent-trait linkages, are certainly more sophisticated and can be chosen as circumstances require. The latent-trait linkages, for example, might allow for the equating of an NRT to an entire pool of test items, hence making it possible to have NRT scores for various custom CRTs. However, unless you have other needs for latent-trait scaling, the effort is probably too great just for equating.

There are some problems associated with the standard equipercentile equating method as proposed by the Title I system. In October 1978, USOE called a subcommittee of professional staff from the Title I Evaluation Technical Assistance Centers to look at the recommended Title I equating procedure for Model A. What follows is a list of the key recommendations from that subcommittee:

1. A hierarchy of strategies for equating should be offered that will help LEAs of varying size and technical sophistication.
2. Technical assistance should be provided to help LEAs:
   a. select an equating strategy.
   b. select an NRT with appropriate score range.
   c. select an NRT with appropriate score range.
3. Investigations should be conducted regarding the potential of latent-trait models for equating tests.
4. The term "equating" might be inappropriate since the CRT and NRT are not parallel tests; the term "estimating" was offered for consideration.
5. Investigations should be conducted regarding the effect of size of equating samples on stable estimates; how small for valid local interpretations or for national aggregation?
6. Equating (estimating) at the time of posttest should be avoided, if possible, since the treatment effect is confounded in the posttest score distributions.
7. Investigations should be conducted regarding the level of correlation necessary between the CRT and NRT.

In addition to that list, another investigation is needed that would assist those wanting to use equated norms with CRTs; if we equate once, how long does that equating last? The Title I system implies that the shortened equating must be done each and every time we evaluate. Is that better than using the results from a more sophisticated equating study for several years?

In an AERA paper, Fishbein (1978) commented on the uses of equating and pointed out, among other things, that the NRT and CRT are not parallel tests. He also reminded us of some of the hazards that Tallmadge (1976) listed for CRT users, especially the need for sufficient score variance. Fishbein also added that while CRTs might be more appropriate in terms of local instructional relevance, they may be so narrowly defined that the resulting evaluations would lose their generalizability to the overall content area. He suggests that an occasional NRT evaluation would be

helpful in determining whether or not the CRT evaluation was leading you astray. This suggestion has excellent merit where there is low turnover, and would help those programs trying to obtain measures of sustained effects.

Some Notions About the Future

The need for CRT and NRT information in the future seems certain--parents, students, teachers, school administrators, school boards and the community as a whole will want to know how well things are going with those things being taught. And, they will want to know if the things being taught and learned are keeping pace with an ever-changing community, nation, and world.

The computer probably holds one of the important keys to the use of CRTs and NRTs. For example, it is entirely possible that a teacher will be able to index a hand-held computer to test specific objectives relevant to a given student. The choice of objectives will be based on performance and instructional information already stored for computer access. The student would then take the test using the same hand-held computer; and the teacher would analyze the results of the test, again using the same computer. It would list performance by objective and by the total set of objectives. It would compare performance to the student's past performance, both in terms of mastery of those particular objectives and in terms of overall progress in the content area. Based on applications of latent-trait models, or the like, the computer would be able to pinpoint fairly accurately where the student was in terms of a single scale for the content area. It would also tell the teacher how well that performance compares with other students, school-wide, state-wide, or nation-wide. The computer would also list the errors made by the student, hopefully as an assist to the teacher or aide for diagnostic-remedial purposes. The computer would store the student's data to form the basis for continued evaluation of various program levels.

Today CRTs and NRTs seem to be separate instruments, which can be merged by administering both to the same students and by looking at results from each. In the future, they may not be separate instruments. Modern measurement and computer technology will be able to merge them into a single continuous evaluation process; and such things as equating will be built into the features of the computer processing of each student's performance on each test item.

One last note seems necessary. Computers often conjure up nasty visions of a mechanistic, inhuman world. The instruction and evaluation of the future cannot avoid the opportunities of technological advances; however, they should be employed to create more time for human interaction and human pursuits. The computer-assisted testing picture predicted above cannot work unless it frees the teacher to merge the results with all aspects of his or her teaching effort and to allow more time for teaching those things only a person can accomplish.

References

Angoff, W. H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Fishbein, R. L. The use of non-normed tests in the ESEA Title I evaluation and reporting system: some technical and policy issues. A paper presented at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Hambleton, R. K., Swaminathau, H., Algina, J. and Coulson, D. B Criterion-referenced testing and measurement: a review of technical issues and developments. Review of Educational Research, Winter 1978, 48, 1, I-47.

Tallmadge, G. K. Criterion-referenced tests: ESEA Title I evaluation and reporting system. Mountain View, California: RMC Research Corporation, October, 1976. (RMC Technical Paper No. 11)