Two Approaches to Criterion-Referenced Program Assessment

TOM HALADYNA

School districts need valid and reliable test data reflecting achievement outcomes of their instructional programs, but a large number rely almost exclusively on standardized tests, which frequently are not directly relevant to local goals and objectives. They can improve program evaluation by using one of two approaches to criterion-referenced assessment: *random sampling* or *item response theory*. Either will provide high quality diagnostic data at low cost and with a minimum of student testing time.

No matter which approach is taken, a series of logical steps will make best use of staff members' talents and their knowledge of the local instructional program.

Step 1: Defining the curriculum. The first step involves defining the curriculum in terms of a set of objectives reflecting both scope and sequence. If, for example, the curriculum involves reading for grades 1–6, the objectives should describe each and every major behavior students must acquire to become satisfactory readers by the district's standards.

Many states, including California, Florida, Michigan, Minnesota, Oregon, and Washington, have developed objectives as the basis for statewide assessment and have provided material for defining local district objectives.

This work was partially supported through a grant from the National Institute of Education. Opinions expressed are solely those of the author and not NIE. The author wishes to acknowledge the constructive comments of Glen Fielding of Teaching Research on earlier drafts of this paper.

Districts looking for better test data may want to use random sampling or item response theory.

Step 2: Identifying or creating items. A recent study (Roid and others, 1980) found that teachers create better test items than professional test makers. That should not be surprising; teachers are more aware of the intent of instruction and can gear both language and content of test items more appropriately to objectives. In addition, there is an emerging technology of test item writing that promises to provide easy, automated generation of many test items (Roid and Haladyna, 1981).

However, a better strategy than writing original test items is to collect and review items from various sources. One such source is the Northwest Evaluation Association (NWEA),[1] a consortium of school districts and other agencies who work cooperatively in the states of Oregon and Washington to improve evaluation in the schools. A major accomplishment of the NWEA is the development of item banks, one in reading, one in mathematics, and one in language usage. Each item bank has been field tested and keyed to instructional goals. Another source of items is the basic skills collection of the Northwest Regional Educational Laboratory (NWREL). Their collection in language arts and mathematics encompasses over 20,000 items that are objective-referenced.[2] One limitation of this collection, however, is the fact that these items have not been carefully field tested and subjected to item analysis and Rasch calibration like the NWEA item bank; nevertheless, the existence of such a resource is impressive to those committed to CR assessment.

Step 3: Item review. Two kinds of item review are recommended: logical and empirical. The first involves examination of items by teachers and other school district personnel to determine if they are appropriate to curricular objectives and grade levels. Once the items have been categorized and keyed to district objectives, the process may take only a single, intensive day of work by a curriculum committee.

In the empirical item review, items are field tested to determine if they "work." We have developed simple, nonstatistical guidelines for evaluating items (Haladyna and Roid, 1981a). Field testing can be a costly process, but it may not be needed if the items have been field tested previously.

Upon completion of these three steps, the district is ready for test design, administration, scoring, analysis, and reporting. Each of the two criterion-referenced assessment plans follows a different path leading to reporting of results, and each has particular strengths and weaknesses.

Tom Haladyna is Research Professor, Teaching Research, Oregon State System of Higher Education, Monmouth, Oregon.

The Random Sampling Plan

Random sampling is the simplest to understand and employ. It has its roots in several theories of testing including classical theory (Nunnally, 1967; Lord and Novick, 1968) and generalizability theory (Cronbach and others, 1972). Most accounts of CR testing (Millman, 1974; Brennan and Kane, 1977; Hambleton and others, 1978; Popham, 1978) recommend that items be randomly sampled from well-defined domains.

Test design involves planning and constructing test forms for assessment. Perhaps the best way to describe the use of test design is by example. Imagine a situation where reading must be assessed for grades two through six following an objective-based curriculum containing 45 objectives considered to be most important to each child's development in reading. Let us also assume that these 45 objectives can be classified into two distinct areas within a general domain of reading: comprehension and word attack skills.

We desire information about academic growth for each of these two areas, as well as overall achievement, at each grade level. We also desire diagnostic information relative to each of the 45 objectives. These objectives do not comprehensively span the curriculum but are introduced in one grade, developed fully in another, and reinforced in successive grades. Therefore, testing should provide information at each of the stages of the developmental sequence as illustrated in Figure 1.

Following the development of test items linked to these objectives through teacher judgments, a pool of items is created that varies from 20 to 50 items for each objective. This pool of items could number from 900 to over 2,000 items across all objectives. This certainly seems imposing, but it is a healthy test-item collection and one that could be useful to a district for many years.

The sampling plan calls for the administration of short test forms to students in the district following a random sampling item selection strategy. What this means is that each child in grades two through six takes a short test, consisting of about 40 items or less, in the fall and in the spring.

An important limitation of the random sampling plan is that it does not make available individual student achievement scores. This is a sacrifice, but the saving is that student testing time is quite short, often less than 30 minutes. If individual student scores are desirable, items left over from the item pool may be used to construct teacher-made mastery tests over specific objectives for classroom use. For instance, a teacher at grade level four could have all mastery tests needed to test for learning of that objective, whether the objective is introduced, developed, or reinforced.

The next step is preparing a testing plan. This district has about ten classrooms at each grade level—a total of 50 classrooms and about 1,250 students—so our testing plan must provide for creation of a sufficient number of test forms. First, a table is constructed representing the number of objectives introduced, developed, or reinforced at each grade level and the corresponding numbers of available items. The number of students is also shown (Figure 2).

Next we start to develop test forms using a random sampling strategy. For example, grade three has 37 objectives and 1,521 items. Since we have 231 students, we might develop five test forms with about 37 items on each form (one item for each objective). Which items to use is indicated in the sampling plan. First, we arrange items by objectives; since there are 37 objectives, we ran-

Figure 1. Developmental Sequence for Four Hypothetical Objectives in a Reading Curriculum.

	Grade					
Objective	1	2	3	4	5	6
1	I	D	R			
2		I	D	R		
3			I	D	R	
4			I	D	R	R

I = Introduce D = Develop R = Reinforce

Figure 2. Items and Objectives by Grade Level.

Grade	Items	Objectives I, D, or R	Students
1	494	12	220
2	960	24	219
3	1,521	37	231
4	1,611	36	216
5	940	24	214
6	541	14	232

Figure 3. Example of a Criterion-Referenced Test Report Based on Random Sampling.

Grade 3	Fall	Spring	Gain
Total (37 Objectives)	48%	63%	15%
Comprehension	52%	70%	18%
Word Attack	44%	56%	12%
Objective			
13	32%	63%	31%
14	44%	46%	2%
15	54%	56%	2%
16	22%	74%	52%

domly select five items for each objective and assign one item to each test form. Our five test forms contain an item representing each objective, and so we have a total of five items for each objective, each form having 37 items. These five test forms are randomly distributed again in the spring so that no student receives the same form. As a consequence, we will have estimates of student performance for each test, each objective, for any combination of objectives, and for all objectives for both fall and spring testing. Because sampling was random, we have unbiased estimates for each reporting unit, for example, comprehension, Objective 23, reading. The unused items can be released to teachers for classroom testing.

The district can increase the coverage of objectives by making tests longer and increasing the testing time, which improves the quality of information. But the option is always present to go with the cost-effective short form, 37 items per form, or the slightly more expensive longer form, which can be 50 or more items. One of the most attractive features of this plan is the resultant report form. For example, Figure 3 illustrates a test report for district administrators, the school board, curriculum personnel, other administrators, parents, teachers, and students.

It should be clear that any and all can understand and interpret such reports because everything is reported in the simple language of a percentage scale. For example, for Objective 14, student performance over the year improved only two percentage points. We might want to consider that information in light of whether the objective was supposed to have been introduced, developed, or reinforced. Not only is the language of the data simple, but it is highly accurate and domain-based.

To summarize briefly the sampling approach:

1. Items are categorized according to whether they are introduced, developed, or reinforced at a given grade level.
2. Tests are designed based on district characteristics and the desired test length and degree of coverage of the curriculum objectives.
3. Items are randomly sampled from the item pool based on the district's information needs.
4. Student scores are not derived, but information is derived for each reporting unit (such as comprehension, Objective 14).

Item Response Theory

During the past decade there has been sudden growth in interest, research, and application of item response theory. The theory fundamentally assumes that student performance on any item is based on the student's level of achievement and the inherent difficulty of the item. The model is statistically sophisticated (Wright and Stone, 1979; Lord, 1980) and provides suitable treatment at that level. It will only be outlined here.

The major advantage of item response theory is that it provides for tremendous flexibility and precision of testing across many grade levels. For example, once an item pool has been field tested and difficulty estimated for each item, one can take any set of items and administer it to any student. The resultant test score can be placed on a single test scale, which underlies all test items. To compare students who took different tests, simply convert their scores to this single test scale.

Thus, you can have group and individual scores that are extremely precise, and tests can be constructed and used with a minimum of effort of test design. At any achievement level, you could employ a single test designed to measure the entire domain of objectives or any subset of objectives.

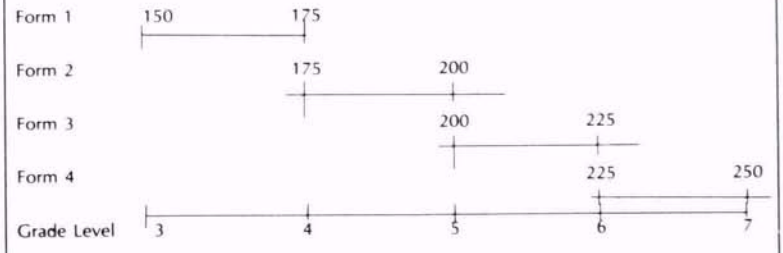
To implement item response theory, we begin with the state of affairs described earlier, with a developmental sequence for each objective and items and objectives by grade level, as shown in Figure 2. We must specify what kinds of information we desire and have compatible field test information.

However, field tests are expensive. If you are fortunate enough to have a field-test item bank, no field testing is necessary. But if you must field test, cooperative interdistrict efforts can be very cost effective and have been successfully done in Oregon and Washington with the cooperation of state departments of education and regional service districts. Others have chosen to join consortia, where interdistrict cooperative and pooled resources result in accomplishments that small school districts could never achieve.

The major difference between item response theory and the sampling model is in test design. With the former, items are selected based on their difficulty. For example, item difficulty can be assessed via a scale that ranges from 150 to 250, with 200 being the mid-point, representing average fifth-grade performance. For our particular example, to test a single objective you might want to develop a test for low performing students with items that ranged in difficulty from 150 to 175. For students in the low range, item difficulty would range from 175 to 200. By careful nonrandom selection of items, you build level tests that comprehensively span the range of performance for the district. Four test forms could span the entire curriculum for any objective, set of objectives, or the entire test-item pool, as shown in Figure 4. The test content depends greatly on the kinds of information you want.

One of the best illustrations of the use of item response theories in program assessment is the Portland (Oregon) Public School District Levels Tests in Reading, Mathematics, and Language Usage. These tests have been administered to nearly 33,000 students in grades three through eight each fall and spring for the past four years. Curricular growth can be measured continuously across these grade levels on a single scale. Other school districts, such as those in Los Angeles and San Jose (California) and Tacoma (Washington), have used similar assessment programs, and statewide assessments have experimented with programs of assessment

Figure 4. Example of the Grade Ranges of Level Tests.



based on item response models. Test companies such as the California Test Bureau of McGraw-Hill, among many others, have been planning new testing programs based on item response theory. Thus, we may expect to see these programs used more widely and successfully in coming years.

Research has shown that when children are confronted with tests that are too hard or too easy, the test results are tremendously biased (Haladyna and Roid, 1981a, 1981b; Holmes, 1980). In fact, those performing Title I assessments are probably faced with this problem and do not have the wherewithal to combat this bias with conventional testing methods. If a child who is functioning at a level of 162 (about the midpoint of form one) takes a form three test, the items are too hard and the score is negatively biased. Positive bias occurs when a student takes a considerably easier test. Therefore, one of the major problems is to assign to each student a test geared to that student's level of functioning. Regardless of the test given, all scores are reported on the common scale, which in this example ranges from 150 to 250 as shown in Figure 5.

Figure 5 illustrates some of the advantages of item response theory in program assessment. For instance, in Title I evaluation, achievement testing must show the progress of Title I students on a fall-to-spring basis. The tabled results indicate this. Holmes (1980) showed how such scores can be converted to normal curve equivalents for Title I reports. In general, it is quite easy to trace the progress of groups of children in a curriculum on this general curriculum scale for any unit of analysis; for example, the total reading domain, comprehension or word attack skills, or a single objective or group of objectives.

Testing for Improved Instruction

These two plans have many advantages over standardized testing and, fortunately, the technology exists to implement them. As districts grow more comfortable with these approaches, we should see an increase in their use and an improvement in the quality of instructional program assessment. High or quality achievement testing will help districts focus more specifically on instructional problems and their resolution, thereby making the testing program a major contributor to improved instruction. ■

References

Brennan, R. L., and Kane, M. I. "An Index of Dependability for Mastery Tests." *Journal of Educational Measurement* 14 (1977): 277–289.

Cronbach, L. J.; Gleser, G. C.; Nanda, H.; and Rajaratnam, N. *The Dependability of Behavioral Measurements*. New York: John Wiley, 1972.

Haladyna, T., and Roid, G. "The Role of Instructional Sensitivity in the Empirical Review of Criterion-Referenced Test Items." *Journal of Educational Measurement* 18 (1981a): 39–53.

Haladyna, T., and Roid, G. "Two Approaches to the Construction of Criterion-Referenced Achievement Tests." Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981b.

Hambleton, R. K.; Swaminathan, H.; Algina, J.; and Coulson, D. B. "Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments." *Review of Educational Research* 48 (1978): 1–47.

Holmes, S. E. *ESEA Title I Evaluation and Reporting Refinement: The Title I Linking Project. Final Report*. Salem, Ore.: Educational Program Audit Division, Oregon Department of Education, 1980.

Lord, F. M. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, N.J.: Erlbaum, 1980.

Lord, F. M., and Novick, M. R. *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison Wesley, 1968.

Millman, J. "Sampling Plans for Domain-Referenced Tests." *Educational Technology* 14 (1974): 17–21.

Nunnally, J. *Psychometric Theory*. New York: McGraw-Hill, 1967.

Popham, W. J. *Criterion-Referenced Measurement*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1978.

Roid, G., and Haladyna, T. *A Technology for Test-Item Writing*. New York: Academic Press, 1981.

Roid, G.; Haladyna, T.; and Shaughnessy, J. "A Comparison of Item-Writing Methods for Criterion-Referenced Testing." Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980.

Wright, B. D., and Stone, M. H. *Best Test Design*. Chicago: Mesa Press, 1979.

[1]For more information, write to Dr. Fred Forster, Executive Secretary, NWEA, Portland Public Schools, 501 N.E. Dixon, Portland, Oregon 97227.

[2]For more information, write to NWREL, 300 S.W. 6th Avenue, Portland, Oregon 97204.

Figure 5. Example of a CR Test Report Form Based on Application of Item Response Theory.

Content	Grade Two Fall	Grade Two Spring	Grade Three Fall	Grade Three Spring	Grade Four Fall	Grade Four Spring	Grade Five Fall	Grade Five Spring	Grade Six Fall	Grade Six Spring
Reading	154	162	160	173	170	194	190	213	209	224
Comprehension	147	158	155	171	167	186	186	210	207	222
Word Attack	161	166	165	175	173	202	194	216	211	226
						I		D		R
Objective 23	—	—	—	—	170	178	174	208	210	224
Objective 25	—	—	—	—	200	212	210	231	232	238