DOCUMENT RESUME

ED 109 239 TM 004 701

AUTHOR Olson, Margot A.
TITLE Criterion-Referenced Reading Assessment in a Large City School District.
PUB DATE 75
NOTE 16p.; 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 Cloze Procedure; *Criterion Referenced Tests; *Educational Assessment; Elementary Secondary Education; Item Analysis; Reading Comprehension; Reading Skills; *Reading Tests; School Districts; Student Evaluation; *Test Construction; Testing Programs; Urban Schools
IDENTIFIERS *Dallas Independent School District; Paragraph Reading Test; Survey of Reading Skills

ABSTRACT

The test-development process used to develop criterion-referenced tests in reading for a large-city school district's system-wide testing program is described. Two types of instruments were developed: (a) a measure of students' ability to read text-books at their assigned grade levels and (b) tests of enabling skills emphasized in the basal reading program. Item construction and selection techniques are described as well as the strategy underlying the use of the two types of tests. The paper is especially relevant to situations where criterion-referenced testing is applied outside the context of instructional development models. (Author)

CRITERION-REFERENCED READING ASSESSMENT IN A LARGE-CITY SCHOOL DISTRICT

Margot A. Olson

Dallas Independent School Distr ct*

Abstract

The test-development process used to develop criterion-referenced tests in reading for a large-city school district's system-wide testing program is described. Two types of instruments were developed: (a) a measure of students' ability to read textbooks at their assigned grade levels and (b) tests of enabling skills emphasized in the basal reading program. Item construction and selection techniques are described as well as the strategy underlying the use of the two types of tests. The paper is especially relevant to situations where criterion-referenced testing is applied outside the context of instructional development models.

A paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., March 31 - April 3, 1975.

*The author is currently with the Learning Resources Center, Eastfield College, Dallas County Community College District.

CRITERION-REFERENCED READING ASSESSMENT IN A LARGE-CITY SCHOOL DISTRICT

Margot A. Olson

In recent years, much emphasis has been placed upon facilitation of learning of basic reading skills in the Dallas Independent School District. Numerous special programs and staff development seminars have provided techniques and materials to aid in the acquisition of basic reading skills. In line with such concern over basic reading skills, a need for competency-based in addition to normative evaluations of reading emerged. To fulfill this need, the Board of Education allocated funds for a two-year project to develop criterion-referenced tests for assessment of reading skills.

The Board of Education approved proposals submitted by the Department of Research and Evaluation to develop two instruments for criterion-referenced assessments: (a) the Paragraph Reading Test (PRT) and (b) the Survey of Reading Skills (SRS). The purpose of the PRT is to assess students' overall ability to gain information from reading materials representing progressive levels of difficulty. The purpose of the SRS is to assess students' acquisition of specific, enabling skills in reading. More specifically, a decision concerning whether students are able to read materials at a reading difficulty pre-determined as desirable can be made using the PRT. Specific skills upon which students would possibly benefit from further instruction can be identified by the SRS. Using both tests, the recommended strategy for testing involves administration of an appropriate level of the SRS to those students who fail to achieve the specified criterion for their grade level on the PRT.

Three persons, a coordinator and two assistants, were assigned to the criterion-referenced test development staff. The coordinator was a specialist in educational measurement, one assistant was a recent graduate in secondary English education, and the other was an experienced, elementary reading teacher. Together these individuals were responsible for development of the reading tests with the understanding that the measurement specialist was primarily responsible for the methodological considerations of development and the reading teacher, for the accuracy and relevancy of the content of the tests.

A brief description of the rationale and process used to develop the PRT and SRS is provided in this paper. The description provides an example of the use of criterion-referenced testing as a part of a large-scale assessment program. Since the literature on criterion-referenced testing (c.f., Cox & Vargas, 1966; Glaser & Nitko, 1972;

Hambleton, 1974) deals with it primarily as a part of instruction and test development within that context, the presentation of test development procedures within the context of a system-wide assessment is especially relevant for others considering wider applications of criterion-referenced testing.

Paragraph Reading Test

Rationale

The PRT was designed to measure the following objective:

> Given an adopted textbook in any curriculum area for the grade level at which a student is classified, the student will be able to decode and comprehend the material presented in the text.

There can be little argument that the long-term goal of reading instruction is to provide each individual with the ability to gain information from printed material that he encounters. As a student such material is often in the form of textbooks selected as instructional media in the classroom. For this reason as well as the results of some research that suggested that students are generally more able to read texts than other materials that they encounter (Hansen & Hesse, 1972), the reading difficulty of the texts at each grade level was selected as the minimum level of acceptable reading achievement.

Measurement of students' ability to read materials of various levels of difficulty has often been a concern of educators. Varieties of techniques for measuring reading comprehension have been proposed and studied. Among these are individually administered oral examinations; examinations requiring written response to multiple-choice, completion, or other common types of items; and the cloze technique. From a brief survey of the literature, it is evident that each of these techniques has validity in certain situations. The special properties of the cloze technique, however, made it especially appealing for measurement of District students' ability to gain information from their texts.

Cloze tests are constructed by systematically deleting words from passages of written prose. Generally, passages of approximately 250 words are used, and every fifth word is deleted and replaced by a blank of uniform length. The examinee's task is to provide the missing words; scores are usually based upon the number of correct restorations. An example of an excerpt from a cloze test is the following:

> The boy _____ girl ran down the _____.
>
> When they reached the _____, they decided
>
> to rest _____ the grass.

Modifications of the cloze technique have been used with success. Among these are deletion of words of specific parts of speech and presentation of alternatives, in multiple-choice format, for the missing words. Several investigators (Gallant, 1965; Kirby, 1968) have suggested that the latter modification is more valid for use with children below third or fourth grade. An example of a cloze test with alternatives is the following:

The boy and girl ran {down / up / across / through} the hill. When they reached the bottom, {boys / they / child / it} decided to rest on the {tree / sofa / top / grass}.

Cloze tests have been used successfully as measures of the relative difficulty of reading selections and as criteria of examinee's comprehension of reading selections (c.f., Bormuth, 1968; Coleman, 1962). As a measure of readability, the difficulties of the passages are graded according to the percentage of correct restorations provided by a specific group of students. When comprehension is measured, comprehension scores are assigned to students according to the number of correct restorations. Numerous studies, (c.f., Bormuth, 1968; Gallant, 1965; Kirby, 1968) have shown cloze test scores to be valid when both oral and written exams have been used as criterion measures.

The most obvious method of constructing cloze tests to measure the objective of student ability to read texts at their assigned grade levels would have been to sample passages from all textbooks of interest. Using only one sample from each major text adopted by the district in grades one through twelve for reading, language arts, social sciences, science, and mathematics would have required development of at least 60 cloze tests. The use of alternate texts in some subjects would require additional cloze tests as well as increase administrative problems. Also, new tests would have to be constructed each time a new textbook was adopted. For these reasons, an indirect approach to the measurement problem was developed. That is, ability of students to read their texts was inferred from a smaller collection of cloze tests.

Methodology

Modified cloze tests incorporating alternatives embedded within the passage like those of the earlier example were adapted from King's (1972) Content Referenced Reading Scale to compose the PRT. The use of multiple-choice format not only facilitated scoring of

large numbers of tests but also enabled use of the tests with students in the lower elementary grades. The group of 16 passages selected by King included reading selections of various difficulties appropriate for elementary and secondary students. Each passage included at least 10 sentences and 150 words and the 10 deletions made in each passage were primarily content words. King's scoring procedure was retained for use with District students: (a) number passages according to predicted difficulty; (b) based upon a 70% correct criterion for each passage assign ones and zeros to each examinee; and (c) assign a total score based on the number of the first passage that precedes the second zero and contains a one. King's research provided evidence that students who could read passages of each difficulty level were likely to be able to read all of the easier passages.

Structural analyses (Dale & Chall, 1948; Spache, 1953) were made of a collection of major texts adopted by the District in reading, language arts, mathematics, social studies, and science; 12 of King's passages; and 14 periodicals (e.g., Saturday Review, Newsweek, Sports Illustrated, Dallas Morning News) of interest to the general population. The structural analyses provided measures of average sentence length, average word length, the proportion of different words, and the proportion of hard words in each reading selection. To establish a readability formula incorporating the four structural variables, a representative sample of approximately 250 elementary school students per grade level responded to cloze tests, constructed in the traditional manner, over three samples from each major text adopted for grades one through eight in the five major subject areas. Using the weighted combination of the structural characteristics, difficulty levels of the texts, test passages, and periodicals were estimated. Observed difficulties were also calculated for texts in grades one through eight. The difficulty values for the test passages and reading materials were matched. Thus, if an individual could read a test passage of a certain difficulty, the generalization was made that he could read other materials of similar or easier difficulty levels. A difficulty score for each grade level was computed by averaging the difficulties of the texts for that grade level. Using the grade level at which the District texts were used, criterion scores for the PRT were derived from the equating process.

Furthermore, representative samples of approximately 500 elementary students per grade level and 250 secondary students per grade level responded to a sequence of the PRT passages appropriate for their grade level. Using the data from these administrations, item statistics, internal consistency reliabilities, and percentage of correct restorations were computed and examined for each reading selection of the PRT. On the basis of the item and test characteristics, a few minor recommendations for revising the PRT were made.

Results

The predicted and observed difficulties of the texts averaged across grade level and the predicted difficulties of the PRT passages are presented in Table 1. In examining the table it should be noted that the PRT criterion passage for each grade level is included in the row containing the difficulties of the texts for that grade level. For example the criterion passage for grade four was number five.

Table 1

Predicted Difficulties of Reading Materials

Texts			PRT Materials	
Grade Level	Predicted Difficulty	Observed Difficulty	Passage Number	Predicted Difficulty
			1	.365
1	.276	.306	2	.301
2	.263	.283	3	.289
3	.202	.183	4	.224
4	.174	.182	5	.187
5	.161	.150	6	.158
6	.142	.129	7	.111
7	.101	.110	8	.102
8	.087	.075	9	.071
9	.077	*		
10	.071	*	10	.033
11	.058	*		
12	.052	*		
			11	-.075
			12	-.028

*Data not available

The predicted difficulties for the periodicals ranged from .091 to -.042 with an average across periodicals of .008.

As might be anticipated the process of selecting criterion scores was not a "neat process." The observed difficulties were selected for the matching, thus making the test somewhat easier at the lower grade levels that it would have been if predicted values were used. Using observed difficulties of texts provided good matching with PRT passages at all grade levels except grade three. Because of the dissimilarity of the predicted difficulties at grades three and four, passage number four had more intuitive

appeal as a useable criterion for grade three. Passage 10 was selected for use for all high school grades for two reasons: (a) the variability of difficulties was smaller (of course, the readability formula derived from elementary students might account for this) and (b) there was less consistency in progressions of difficulties of texts within subject matter areas than was evident in lower grades. PRT passages 11 and 12 provided criterion levels appropriate for generalizing to the periodicals or, perhaps, other materials more difficult than the typical secondary text.

The percentage of correct restorations on the 12 PRT passages and the internal consistency reliabilities are presented in Table 2.

Table 2

Percentages of Correct Restorations for the PRT Passages by Grade Level

Grade Level	Passage 1	2	3	4	5	6	7	8	9	10	11	12	Reliability
1	54	41	34	*	*	*	*	*	*	*	*	*	.89
2	74	64	54	44	*	*	*	*	*	*	*	*	.92
3	82	74	63	55	44	*	*	*	*	*	*	*	.93
4	87	79	71	64	50	42	*	*	*	*	*	*	.93
5	91	85	76	71	55	48	46	39	*	*	*	*	.94
6	92	86	78	74	58	54	53	46	48	45	*	*	.96
7	*	91	*	84	67	62	59	51	54	52	38	36	.91
8	*	93	*	86	67	66	64	57	60	58	41	41	.93
9	*	92	*	89	74	71	71	63	63	61	47	42	.95
10	*	96	*	94	79	82	81	74	78	71	55	50	.89
11	*	97	*	94	76	80	78	72	75	70	51	50	.91
12	*	96	*	94	80	84	82	78	79	75	61	55	.94

*Not administered at this grade level.

As anticipated, the difficulty levels of the PRT passages generally increase as one goes across the rows in Table 2. The scalability of the passages as reported by King (1972) was substantiated by the administration to a sample of District students. Item analysis data was used to improve the scalability of the passages by attempting to change the difficulty of some deletions.

The PRT is intended to be administered to all children in grades one through twelve as a part of the system-wide testing program. Results across students are presented in a manner similar to the testing profiles often associated with norm-referenced tests. In addition, the score is entered into the student's permanent record. A short description and table for interpreting the

scores is available to all teachers. The results of the PRT should alert teachers to students who might have reading problems (i.e., fail to read and comprehend texts at their assigned grade levels) and, thus, provide additional information possibly useful in designing appropriate reading instruction. The use of the SRS is recommended as an aid in determining areas of remediation for a student who fails to meet the minimum criterion for his grade level on the PRT.

Survey of Reading Skills

Rationale

To develop a criterion-referenced test for purposes of assessing enabling skills, it is generally recommended that performance objectives, stated in terms of readily observable behaviors, be written to delineate specific skills of interest to the examiner. Until recently such objectives were used primarily in the design of instruction, and tests to accompany the objectives were administered prior and subsequent to instruction. This permitted the examiner to measure every skill covered by the instruction and to space the opportunities for testing across an extended period of time.

As the uses of criterion-referenced tests have been extended over the past few years, deviations from the usual applications of criterion-referenced testing as a part of the curriculum have increased. One increasingly more widespread use of criterion-referenced tests is occurring within the context of large-scale assessment programs. In such situations the number of skills to be assessed must be limited since it is desirable to administer the tests within a relatively short period of time. Because of limited testing time, it is necessary to select a reasonable number of the most relevant objectives for testing purposes.

Within the context of curriculum, test items can be written in the form which permits the most direct method of measurement. It is necessary, however, for large-scale testing purposes, to construct items so that they can be objectively scored. This restricts the item format to those which provide alternative answers. After sufficient numbers of items have been constructed, they must be compiled into test booklets. Within the context of curriculum, a separate test booklet for each skill is appropriate. For large-scale testing purposes, however, where tests over multiple objectives are administered simultaneously, it is more reasonable to combine the tests over each objective into a single survey. The resultant survey of skills might more appropriately be considered a test composed of numerous subtests.

The final phase of a test development process involves a tryout of the test items. Since the purpose of testing is to decide whether an examinee has or has not mastered the skills represented

on the tests, it is important to determine if the items will perform as anticipated. On the basis of an analysis of the item characteristics, items should be selected so that the best estimation of an examinee's mastery of a skill can be made. Generally acceptable items are retained, other items are discarded or revised, and/or new items are written; and a new test is created which supposedly has better technical qualities than the original test.

A procedure much like that just described has been used to construct the SRS series of tests. Performance objectives were written to specify the most relevant skills for composing each level of the SRS series. Items measuring each objective were constructed and combined to form the various tests. The surveys were administered to representative samples of District students in kindergarten through grade twelve for purposes of evaluating the effectiveness of test items. Based upon the results of the tryout data, revised tests were constructed.

Methodology

Test Development. The SRS was developed to measure reading skills emphasized in the District's basal reading program in kindergarten through grade six. Three groupings of tests compose the SRS series: (a) Level K, appropriate for measuring pre-reading skills, (b) Levels I-VI, appropriate for elementary students, and (c) Level S, appropriate for secondary students who fail to learn basic skills during their elementary years. The groupings and specific tests are presented in Table 3.

The Level K tests are appropriate for students who are acquiring the skills needed prior to beginning instruction over basic reading skills. The Level I-VI tests correspond to skills emphasized in the District's basal reading curriculum for the elementary grades. Selection of the appropriate level of the elementary SRS to be administered to individual students should be based not upon the student's grade level but upon the difficulty level of the hardest PRT reading selection successfully read as well as the student's position in the basal reading program. The Level S tests measure cross-sections of basal reading skills in four general areas and contain items of an interest level appropriate for secondary students who have not acquired the skills normally associated with the basal reading program.

To delineate groups of skills relevant for each of the levels of tests, performance objectives were written by the testing staff with the aid of several committees of teachers and administrators. An example of an objective from each of the groupings of tests is given below.

Level K: Given five printed capital letters, two of which are the same, the student will identify the matching letters.

Table 3

Tests Composing the SRS Series

Grouping	Tests
Kindergarten or Pre-reading	Level K--Group Test Level K--Group Manual Level K--Individual Test Level K--Individual Manual
Elementary	Level I Test Level I Manual Level II Test Level II Manual Level III Test Level III Manual Level IV Test Level IV Manual Level V Test Level V Manual Level VI Test Level VI Manual
Secondary	Level S--Phonics Test Level S--Structural Analysis Test Level S--Word Meaning Test Level S--Comprehension Test Level S--Manual

Levels I-VI: Given a picture of a scene, the student will determine whether a printed sentence describes the scene.

Level S: Given a printed sentence with an underlined word, the student will demonstrate ability to derive meaning from context by selecting one of several alternate definitions of the underlined word.

The Level K individual tests were based upon 27 performance objectives; the Level K group test, on 20 objectives. Levels I-VI measured skills expressed by 99 objectives and Level S test measured skills expressed by 40 objectives. Word lists were developed to accompany the various levels of the SRS and more detailed item specifications were written to further clarify the types of items appropriate for each objective.

To provide an easily scorable test, multiple-choice items were written for each skill. Most items included four alternative answers. Sufficient items were included in the tests to insure a high probability that a decision of mastery of a skill was not attributable to chance (Gulliksen, 1950, p.263). No effort was made to provide an estimate of the proportion of items within a specified domain that could be answered correctly by examinees. Since the short tests (or subtests) for each skill were printed together to compose each specific test in the SRS series, an attempt was made to maximize the number of skills measured by using a minimum number of items to measure each skill. Most subtests included four test items. Although four items are generally not sufficient for making highly reliable decisions about individuals, it was decided that many "clues" would be more useful to teachers than a few very reliable subtests composed of many more items.

Twice the desired number of items were constructed for the preliminary forms of Levels I-VI and Level S. Items for each objective were randomly assigned to one of two test forms (Forms A & B). Due to the difficulty of locating stimuli (e.g., cat, dog, house) which would be recognizeable by all District kindergarten students, items sufficient for only one form of the Level K tests were constructed.

Samples of District students were selected for tryout purposes. The sample for the Level K tests included 1174 kindergarten (i.e., five-year olds) students. Approximately 125 elementary students at each grade level responded to either Form A or Form B of the SRS level having the same number as their grade level (e.g., students in grade two responded to the Level II tests). Approximately 250 secondary school students who were assigned to remedial reading classes responded to both Forms A and B of one of the four Level S tests (e.g., a total of 250 students from across all six grade levels responded to the Phonics Test). Prior to testing, the mastery status of the examinees was unknown. It was assumed that a representative sample of students from all parts of the continuum of acquisition of each skill would be involved in the item analysis.

Item Selection Technique. A study of methodologies appropriate for selecting items for criterion-referenced tests over specific enabling skills was investigated as a part of the test development program. Since most techniques for selecting items for criterion-referenced tests require knowledge of the instructional history of the examinees and are based upon pretest-posttest comparisons (c.f., Ivens, 1970; Ozenne, 1971; Popham, 1971), a methodology was needed for application in situations where pretest-posttest comparisons were impossible or impractical to implement. For this reason, the use of an item validity approach for selecting items was investigated. To avoid criticism that criterion-referenced tests often possess no variability, samples of examinees were selected, regardless of their instructional history, from the general population for which the tests were appropriate (Woodson, 1974a; 1974b).

Data collected for the evaluation of Level III of the SRS were used in the investigation of item selection techniques. To estimate item validity, point-biserial correlations were computed using three external criterion variables. Two, the PR and the reading comprehension section of the Comprehensive Tests of Basic Skills, represented measures of the overall domain of reading. The third represented a score over several similar criterion-referenced tests.

Data analyses, across two samples of third-grade students, (i.e., the sample mentioned in th ding section and another drawn in a similar manner) and t of the Level III tests, were performed to compare the va cy indices based upon the three external criterion variables. Comparisions were made across all items as well as within groups of items representative of easy, medium, and hard difficulties and across groups of examineees selected on the basis of expected differences in ability (samples of second graders and fourth graders were selected, also). The item validity indices were compared, also, to techniques based upon an internal criterion variable and sensitivity to instruction. Using third-grade examinees, the internal criterion index represented a difference in item performance between groups assigned mastery and nonmastery status on the basis of their subtest scores. The sensitivity-to-instruction index was obtained by subtracting the proportion of a group of second graders who passed each item from a group of fourth graders who passed the item. As a final comparison, the item validity indices were compared to the traditional item discrimination index based on a total score across all items of each form of the Level III test.

Results

Results of the investigation of item selection techniques (Olson, 1974) revealed that item validity indices obtained using the three external criterion variables were quite similar. In general, there was a high degree of consistency among indices based upon items of different difficulties, among indices based upon two groups of third graders, and among indices based upon groups of second and fourth graders. The high correlations among indices were consistent across both test forms. The strength of the relationships among the indices provided evidence that the item validity index is a viable technique for selecting items for criterion-referenced tests.

Comparisons of the indices based upon item validity, an internal criterion variable, and sensitivity to instructio revealed little similarity among techniques. It is probable that judgments of the item characteristics would be different depending on the technique used. The relationships among the traditional item discrimination index and the item validity indices were strong, however; and, due to the similarity of computational processes used to obtain the traditional index and the item validity index based on similar skills, these two indices could be used interchangeably in situations when responses of an appropriate group of examinees displaying variability over a number of tests of similar skills are available. Thus, the

traditional item discrimination index appeared to be suitable for evaluating the effectiveness of an item in discriminating the mastery status of examinees in a large-scale assessment situation where samples of students demonstrating variability of response exists.

Using the conclusions of the investigation of item selection techniques, data were analyzed for each of the three groupings of tests and items were retained, revised, and occasionally replaced. The resultant tests composing the SRS series are provided to teachers for use on a voluntary basis. The scores provide an indication of which skills students have and have not acquired. They should be useful, as stated earlier, for identifying skills still to be acquired by students who do not achieve the expectation for their grade level on the PRT. Also, the use of the SRS as part of a systematic data collection program should provide valuable information to administrators about reading achievement across specified groups of students.

Conclusions

Further work, of course, is needed on the PRT and SRS to determine the statistical properties of the revised tests and their relationships to each other and to norm-referenced tests used by the District. When an educational agency takes the option of doing its own test development, however, they sometimes sacrifice the very extensive statistical work often associated with published tests for content specifically appropriate for their population. The work presented in this paper provides a good basis for a continuing assessment of reading competency. However, curriculum and needs change and there is a need for constant revision and updating of objectives and test items. Alternate forms of the PRT are possibly needed. Thus, continuing usefulness of the tests will be dependent upon not only some basic research to determine their properties and relationships to other tests but also an avenue to constantly keep their content relevant to the needs of the teaching staff.

References

Bormuth, J. R. The cloze readability procedure. Elementary English, 1968, 45, 429-436.

Coleman, E. B. Improving comprehensibility by shortening sentences. Journal of Applied Psychology, 1962, 46, 131-134.

Cox, R. C. & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1966. ED 010 517

Dale, E. & Chall, J. S. A formula for predicting readability: Instructions. Educational Research Bulletin, 1948, 27, 37-54.

Gallant, R. Use of cloze tests as a measure of readability in the primary grades. In J. A. Figurel (ED.), Reading and inquiry. International Reading Association Conference Proceedings, 1965, 10, 286-287.

Glaser, R. & Nitko, A. Measurement in learning and instruction. In R. L. Thorndike (ED.), Educational measurement. Washington, D. C.: American Council on Education, 1971, 625-670.

Gulliksin, H. Theory of mental tests. New York: John Wiley & Sons, Inc., 1950.

Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. Review of Educational Research, 44, 1974, 371-400.

Hansen, L. H., & Hesse, K. D. An interim report of results of the pilot assessment of reading literacy. Office of Research and Testing, Madison Public Schools, April, 1972.

Ivens, S. H. An investigation of item analysis, reliability, and validity in relation to criterion-referenced tests. (Doctoral dissertation, The Florida State University) Ann Arbor, Mich.: University Microfilms, 1970, No. 71-7036.

King, F. J. Development of a content referenced reading scale. Unpublished manuscript. Florida State University, 1972.

Kirby, C. L. Using the cloze procedure as a testing technique. Paper presented at the International Reading Association, Boston, 1968. ED 019 202

Olson, M. A. A comparison of three techniques for selecting items for criterion-referenced tests. (Doctoral dissertation, The Florida State University) Ann Arbor, Mich.: University Microfilms, 1974.

Ozenne, O. G. Toward an evaluative methodology for criterion-referenced measures. (Doctoral dissertation, University of California, Los Angeles) Ann Arbor, Mich.: University Microfilms, 1971, No. 72-2881.

Popham, W. J. Indices of adequacy for criterion-referenced test items. In W. J. Popham (ED.), Criterion-referenced measurement: An introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971, 79-98.

Spache, G. A new readability formula for primary-grade reading materials. Elementary School Journal, 1953, 53, 410-413.

Woodson, C. E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64. (a)

Woodson, C. E. The issue of item and test variance for criterion-referenced tests: A reply. Journal of Educational Measurement, 1974, 11, 139-140. (b)