DOCUMENT RESUME

ED 186 457 TM 800 146

AUTHOR	Walker, Clinton B: And Others
TITLE	CSE Criterion-Referenced Test Handbook.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	79
CONTRACT	400-76-0029
NOTE	266p.
EDRS PRICE	MF01/PC11 Plus Postage.
DESCRIPTORS	Achievement Tests; Annotated Bibliographies; Cognitive Objectives; *Criterion Referenced Tests; Elementary Secondary Education; *Evaluation Criteria; Resource Materials; *Test Reviews; *Test Selection
IDENTIFIERS	Test Bibliographies

ABSTRACT

The bulk of this document consists of reviews of over 60 criterion-referenced tests, most of which are used to test elementary or secondary-level achievement in the basic skills. For each test review, the following information is given: description of test, price, field test data, administration, scoring, and other comments. The tests are rated according to three categories of criteria: (1) conceptual validity--domain descriptions, agreement, and representativeness; (2) field test validity--sensitivity, item uniformity, divergent validity, lack of bias, and consistency of scores; and (3) appropriateness and usability--clarity of instructions, item review, visible characteristics, ease of responding, informativeness, curriculum cross-referencing, flexibility, alternate form availability, administration, scoring, recordkeeping, decision rules, and comparative or normative data. Guidelines on aspects of test selection are given: locating tests, comparing tests' technical and practical features, and comparing tests for their curricular relevance. Appendices list resources for developing or purchasing criterion-referenced tests, sources of other test reviews, definitions of terms, and available tests which were not reviewed. A subject index to the reviewed tests, a directory of publishers, and a sample of an exemplary domain description are also included. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED186457

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. Young

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CSE
CRITERION-REFERENCED TEST HANDBOOK

IM800 146

CENTER FOR THE STUDY OF EVALUATION

UNIVERSITY OF CALIFORNIA • LOS ANGELES

CSE CRITERION-REFERENCED TEST HANDBOOK

CSE Test Evaluation Project

Project Staff:

Clinton B. Walker, Director
Margeret Dotseth
Russell Hunter
Karen Ogg Smith
Lynnette Kampe
Guy Strickland
Shonagh Neafsey
Christine Garvey
Michael Bastone
Elizabeth Weinberger
Kathi Yohn
Laura Spooner Smith

Center for the Study of Evaluation
University of California, Los Angeles
1979

The Center for the Study of Evaluation is an educational research and development center established in 1966 by the U.S. Department of Health, Education and Welfare. This book and much of CSE work is supported by a contract with the National Institute of Education.

The mission of the Center is to conduct programmatic inquiry in the nature of testing and evaluation in public education.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

The authors are happy to acknowledge the many contributors to this volume. A draft of the "Introduction to the Test Reviews" (Chapter 3) was reviewed during its development by Jason Millman, Professor of Education at Cornell University; Jack C. Merwin, Dean of the College of Education at the University of Minnesota; Albert H. Rouse, Jr., Department of Research and Development, Cincinnati School District; and members of the professional staff of CTB/McGraw-Hill.

Albert H. Rouse, Jr. also suggested some of the basic ideas which were used to develop the procedure for finding the test with the greatest relevance to a given curriculum (Chapter 6). A draft of that procedure was reviewed by Doris Morton, Master Teacher at Hawaiian Avenue Elementary School in Los Angeles; and comments on it were received as well from James Cox, Evaluation Consultant with the Office of the Los Angeles Superintendent of Schools. Lynn Lyons Morris of the Senior Research Staff at CSE made extensive improvements in Chapter 6.

Earlier versions of the text were reviewed by Jason Millman; Carolyn Denham, Associate Professor of Education at California State University, Long Beach; Jeffrey S. Davies, Coordinator of Research, Evaluation, and Testing, Ventura (CA) Unified School District; and Joan Herman and Rand Wilcox of the Senior Research Staff here at CSE. Robert Stake, Professor of Education at the University of Illinois, made critical comments on parts of the text while he was a Visiting Scholar at CSE. Howard Sullivan, Professor of Education at Arizona State University, also gave us many useful comments on the text.

Comments on the text were also received from James Block, Professor of Education at the University of California, Santa Barbara; Thomas Haladyna, Associate Research Professor in the Oregon State System of Higher Education; Joan Bollenbacher, Director of Testing Services, Cincinnati Public Schools; and Thomas J. Riley, Director of Research and Evaluation, Fresno (CA) County Department of Education.

We are deeply grateful for the improvement which each of these reviewers has brought to this volume. The reader should note that the reviewers did not agree with everything herein and that, in places, we did not take their good advice.

Our appreciation goes also to other CSE staff members for their patience, insight, and support. Laura Spooner-Smith and James Burry helped with organization and editing, and Marlene Henerson did extensive final editing. Much of the subject index was done by Diane Ornstein and Laura Spooner-Smith. Correspondence with reviewers and publishers, as well as drafts of the volume, were ably typed by Phyllis Burroughs, Donna Cuvelier, Irene Chow, and Allison Hendrick. Donna Cuvelier has our special thanks for doing the layout, formatting, and typing of the final manuscript.

This project was supported by the National Institute of Education (NIE) under Contract No. 400-76-0029. However, the opinions and findings expressed here do not necessarily reflect the position or policy of NIE, and no official endorsement by NIE should be inferred.

FOREWORD

CSE Criterion-Referenced Test Handbook is the sixth in a series of test evaluation books prepared by the Center for the Study of Evaluation (CSE). CSE is a federally funded research and development center associated with the Graduate School of Education at the University of California, Los Angeles (UCLA). In 1970, CSE published the first book of test reviews, *CSE Elementary School Test Evaluations*. In that volume and subsequent volumes, standardized, norm-referenced tests designed for use in schools were reviewed and rated. The present volume is the first in the series to deal with criterion-referenced tests.

In deciding which tests to review for this volume, CSE staff proceeded in two stages. First, we conducted a wide ranging search for likely tests; we then screened the resulting pool of measures. We examined the catalogs of hundreds of test publishers, bibliographies of tests (listed separately under References), and test lists compiled during previous CSE projects. A retrospective and ongoing search of the Educational Resources Information Center (ERIC) system was conducted using the following subject headings: criterion-referenced, mastery, objectives-based, domain-referenced, content-standard, and universe defined tests. The retrospective search covered the past ten years of *Current Index to Journals in Education* and *Research in Education* and the past five years of *Psychological Abstracts*.

All leads to possible CRTs were pursued by letters of inquiry and, for non-respondents, follow-up letters. The letters of inquiry requested information on tests which fit any of the descriptors used in the ERIC search, which were designed for any of grades K-12, and which were available to test users apart from instructional materials. We then ordered sample materials for each test and later in the course of the project checked with publishers to ensure that we had on hand the most current and complete information to support each test. Seventy-seven commercial publishers and ninety-two non-commercial test developers (mostly school systems) were contacted as possible sources of CRTs. The thoroughness of the search was cross-checked and confirmed by the responses of a national sample of 421 school district staff members who replied to a survey question on CRT use in their districts.

The variety of available measures required that rules be developed for screening tests for inclusion in this volume. Screening

rules were developed both on theoretical grounds and in response to the idiosyncracies of the available measures. The first screen was *availability*: only tests that are readily available to general test users were included. About 80 locally developed tests were excluded when this rule was applied. During the course of the project, some tests were dropped from the list because they were removed from the market. Developmental or experimental versions of tests which were available in only single copy were not included, since such tests sometimes do not go into production or, when they are produced, often appear in a form quite different from the developmental version. Also excluded were three tests of a publisher which required the prospective buyer to visit the sales location.

The next screen resulted from our working definition of the concept of CRT. Since a technically strict definition would have excluded all of the available measures, a less stringent definition was used. The need to acquaint test users with the current set of approximation to CRTs dictated using the following four part definition:

- The measure was originally designed to indicate an absolute rather than a normative level of learning.
- The measure was built around explicit objectives.
- The test items are keyed to these objectives.
- Scores are provided for each objective.

The first part of the definition excluded tests originally developed as norm-referenced tests to which objectives were later added. Also excluded were tests of typical performance such as attitude tests. Measures which did not meet this or other screens discussed below are listed in Appendix E.

Among the tests that were readily available, only those that were not embedded in a special curriculum were reviewed. This rule excluded tests which are sold only with curricular materials or which, although sold separately, are keyed to the content and organization of one curriculum. This rule was adopted since such tests are acquired mainly as a result of a decision about teaching materials. Our system for evaluating CRTs, described in Chapter 3, may still be applied locally to such tests as a part of the process of choosing among curricular series.

Another class of readily available tests that were not reviewed were the customized or made-to-order CRTs which a few publishers offer. Test users with sufficient funds could probably hire any test publisher or consulting firm to create CRTs for a specific curriculum. A listing of publishers who offer this service routinely is given in Appendix A.

Some possible CRTs were excluded on other grounds. Tests that would have to be duplicated by a photocopying method were screened out. Of those, the materials that are uncopyrighted are listed in Appendix A as Resources for Developing CRTs Locally. Behavior checklists were excluded (e.g., Can the child tie his shoes? Can the child skip?). Measures of behaviors that are usually the result of maturation or general experience were also excluded. Finally, tests with only one item per objective were excluded on the ground that they were not serious attempts at criterion-referenced measurement. Two exceptions to this rule were made owing to the likely attention these tests will receive as a result of extensive publisher promotion.

While the acquisition and screening of tests were taking place, project staff developed a set of standards for evaluating CRTs. Although some possible test features are not relevant in all tests or for all test uses, the need for test users to be able to compare tests dictated the development of one evaluative scheme for use across tests. An initial pool of 70 test features was developed on the basis of a review of the professional literature and test publishers' promotional materials. This large number was reduced by several methods. First, some judgments were combined, for example, test-retest and alternate form reliability. Next, features which were more relevant to NRTs than to CRTs were eliminated. Finally, test features which could only be evaluated with respect to a local testing situation (e.g., estimated time for test administration) were treated as descriptive rather than evaluative information.

A draft of the evaluative system was reviewed externally by authorities in CRM and then tried out on a sample of tests. The final version of the system, given in Chapter 3, reflects these reviews as well as the input from a national survey conducted by CSE of 530 test users on school district staffs. The system was also sent for comment to test developers whose products were being screened for this volume. Only two of the test publishers replied.

Each test was evaluated independently by two members of the project staff. These staff members were beyond the M.A. level in

education and had extensive experience on previous test evaluation projects, in test use, and in evaluation. All evaluations were reviewed by a third judge who adjudicated any differences between the original evaluations. The project director then checked and edited all the test evaluations. This process resulted in a rate of agreement in evaluative judgments of 88.5%.

For all of the tests that survived screening, the complete reviews were sent in March, 1978, to the test developers for comment. In some cases, the test developers provided information that persuaded us to change some aspect of our review. In other cases, we were not persuaded by the publishers' feedback, but we report it with the test review. In all, sixteen publishers replied and twelve did not.

In the course of searching for CRTs, we unearthed a variety of resources which are potentially useful to the readers of this volume. These resources are described in the appendices.

Before final editing, the text was reviewed by school and district level educators, university faculty in education, and evaluators at CSE who are noted in the Acknowledgments.

CHAPTER I
Introduction

This chapter provides a summary of the book's contents and makes suggestions for using the various parts of the book according to the reader's specific needs.

Testing influences our lives in many ways. When we were children, our grades in school, course of study, access to higher education, and even self-image were determined in part by our performance on tests. As adults we relive many of those experiences through the children in our lives. With our taxes, we pay for the education of children; and when the test results of educational programs are made public, we are consumers of the scores.

A growing awareness of the impact of testing has caused educators and researchers to look more critically at existing measurement tools. In particular, standardized norm-referenced tests--their social fairness, sensitivity to students' learning, and relevance to instructional decision making--have come under attack, leading in the extreme case to calls for a moratorium on testing in schools.

Some critics of these tests, in their search for more constructive remedies, have turned to the technology of programmed instruction. A major component of programmed instruction is frequent testing of small units of study. This approach to testing is seen to hold promise for meeting some of the major objections to the conventional methods of measurement. Since the test items in the programmed materials use the concepts and content of instruction, they have diagnostic usefulness. Their sensitivity to learning of the materials seemingly reduces their sensitivity to students' social backgrounds. They are thus seen as less biased, more "culture fair."

Criterion-referenced testing[1] (CRT) is partly an outgrowth of this technology. As educators have come to recognize that testing, evaluation, or indeed all of educational management should better support the continuing renewal of instruction, the appeal of instructionally relevant tests has grown. Major test publishers have developed and marketed criterion-referenced tests, and nearly half of the school districts in the United States now report using such measures.[2] The *CSE Criterion-Referenced Test Handbook* was undertaken in response to these developments in educational measurement.

[1]A glossary can be found in Appendix C.

[2]Dotseth, et al., 1978.

The Contents and Uses of This Book

The *Criterion-Referenced Test Handbook* is a collection of resources for educators who develop testing programs, use tests, or need merely to stay informed about advances in educational measurement. The work for this book was driven by two beliefs:

- That testing should support instruction as directly as possible, and
- That source materials on testing should be easy for test users to apply.

This volume is meant to function as an introduction to criterion-referenced testing and as a guide for selecting tests. Readers who want an introduction to the basic concepts of CRT can start with Chapter 2, which contrasts CRT with the standardized, norm-referenced approach, and proceed to Chapter 3, which introduces the test reviews. A framework for evaluating criterion-referenced tests is given here which describes the importance of 21 test characteristics. Basic sources are listed in the References for those who would read further on the subject of criterion-referenced testing. A Glossary designed to explain basic evaluation and measurement concepts in a non-technical manner is provided in Appendix C.

To survey the nature and quality of available CRTs, readers may refer to the evaluative and descriptive reviews that make up Chapter 4. Test selection can also begin here. Identification of likely tests starts by referring to these reviews which are indexed at the back of the book by test name (Index A), test content (Index B), and publisher's name (Index C). Index C also includes publishers' addresses for ordering the current year's test catalogs while Appendix B lists sources of other test reviews.

Secondary sources, such as test reviews and publishers' catalogs, are not sufficient, however, to tell which of several seemingly appropriate tests is best for a particular pupil population, curriculum, and testing need. To make such a choice effectively, test buyers need to study the different tests directly. Chapters 5 and 6 give step-by-step procedures for comparing tests first hand. In addition to giving an overview of the process of test selection, Chapter 5 guides the reader in comparing tests' practical and technical merits for the given testing situation. The single most important feature of tests, their specific relevance to the local curriculum, is finally evaluated by methods which are detailed in Chapter 6.

The guidelines in these last two chapters have much broader application than just to the tests reviewed in Chapter 4. They can be applied to any achievement tests, CRT or NRT (norm-referenced test), reviewed or not reviewed.

If no suitable tests emerge from the steps in Chapter 5 and 6, or if the reader begins with the intent to develop criterion-referenced tests locally, the references to item banks and test development guides in Appendix A will be helpful.

CHAPTER 2
Basic Concepts and Issues

This chapter introduces criterion-referenced testing by comparing it with standardized, norm-referenced testing. The points of contrast are the form and meaning of test scores, the methods used in developing the test, and the optimal test uses. The importance of curricular relevance in testing is stressed. The chapter concludes with a discussion of issues in criterion-referenced testing.

The Form and Meaning of Test Scores

Criterion-referenced testing (CRT) is informally contrasted with norm-referenced testing (NRT) in these terms: criterion-referenced tests (CRTs) are said to show *what a person knows* or *can do*, while norm-referenced tests (NRTs) show *where a person ranks* in a group of test takers. CRTs indicate how completely the student has learned a skill or body of information, while NRTs show where the student stands in comparison with other students--that is, compared to a norm group.

This informal contrast captures an essential difference between NRT and CRT, namely, how the test scores are interpreted. Scores on CRTs are based on a scale of 0 to 100% correct and are indicators of the test taker's thoroughness or completeness of learning (or knowledge, skill, or competency) in the domain being tested. Thus, CRT scores are supposed to be directly meaningful in terms of the degree of learning which the individual test taker possesses. Scores of other test takers do not enter into the criterion-referenced meaning of test results.

NRTs, on the other hand, yield raw scores which are converted to percentiles, grade equivalents, stanines, or other numbers referring to where a score stands among the scores of other test takers. Norm-referenced meaning tells how well one student did in comparison with a norm group; criterion-referenced meaning tells how well a student did compared with how well it is possible to do.

The difference in meaning between the two types of test scores is like the difference between a runner's *time* for finishing a race--a criterion-referenced test--and that runner's *place* (first, second, etc.) --a norm-referenced test. The two types of results are meaningful, but they give different information. In some cases, the two types of results will give a different impression of the same performance. For example,

in a field like physics or gymnastics, where most people have limited knowledge or skill, one might score in the very high percentiles on a test of the general population while being far from learned or skillful. Conversely, in a skill where many people are well trained, such as driving a car, a very proficient test performance might earn a norm-referenced score in only the middle percentiles of the general driving public.

Current writings on testing abound with problems of terminology. To begin with, dictionaries do not recognize the word *reference* as a verb. Authors use the terms *criterion, criterion-referenced, domain-referenced, norm-referenced*, and *standardized* differently. Worse still, these authors rarely make clear whether they mean to reflect common usage or to improve it.

The term *criterion*, for example, is often used to mean *cut score* or *lowest acceptable score*. In this context, a CRT is a test with such a cutting score, where results are treated in pass-fail terms. Elsewhere the term *criterion* is defined as *the specific skill or knowledge being measured* and is used interchangeably with the term *domain*. The domain/criterion can be viewed as the larger, perhaps unlimited, set of potential test items from which the actual test items are drawn. In this context, a CRT is a test that gives *domain estimates*; that is, a CRT estimates the proportion of the domain which the test taker knows or can do. In this book, the terms *criterion* and *criterion-referenced* have the latter meaning.

Some authors use the term *standardized tests* to refer to any ready-made (or off-the-shelf) published tests. Others use the term to mean *norm-referenced tests* (NRTs). The term standardized test will be used in this book to refer to NRTs.

Methods Used in Developing the Tests

In principle, there can be both norm- and criterion-referenced meaning for scores on a single set of test questions. A number of tests reviewed in this book give both norms and absolute scores. There is reason to believe, however, that a test which is most effective for rank-ordering students is less effective as a direct index of their proficiency, and vice versa. The methods used in test development determine whether a set of questions will function better to locate pupils on a scale of learning or on a scale of other test takers. The two main differences in test development are these: how fully the behaviors to be tested are described, and how items are chosen to be on the test.

First, NRTs are generally designed to measure such broad educational goals as "reading comprehension" or "word attack skills." Theoretically, the behaviors to be tested on a CRT are described in much greater detail. The specifications for a CRT are, in theory, detailed enough to describe the content and format of all possible items on the test, thus describing the scale of learning which the test measures. A test is effective in indicating the degree of learning (i.e., in functioning as a CRT) only to the extent that it provides a clear description of what is to be learned. In current practice, few of the tests which are sold as CRTs report using such detailed test specifications. As a group, the existing CRTs achieve a clearer description

of the behaviors to be tested than NRTs typically do in that they break down the broad educational goals into more specific skills. For example, on a CRT, reading comprehension may be divided into literal and interpretive comprehension; then interpretive comprehension may be further divided into separate tests dealing with cause and effect, paraphrasing, real vs. make believe, fact vs. opinion, relevance of statements, stated vs. unstated assumptions, analogy, predictions, and so on.

A second difference in test development is the way in which items are screened for inclusion on the test. Since a CRT is supposed to reveal the thoroughness of learning, it performs that function best when the test items are a representative sample of the material to be learned. Test items are thus selected for a CRT on the basis of whether they are congruent with the test's specifications, that is, the detailed description of the test's content and format.

NRTs, on the other hand, are intended mainly to discriminate or rank individuals; and items which do this best are selected for inclusion. A test gives the most consistent ranking when it produces a wide range of scores. Test items which are very easy do not help to differentiate test takers because everyone tends to get them right. Similarly, very hard items do not discriminate among test takers, for everyone tends to miss them. In order to produce the greatest and most consistent differences among people's test scores, one selects items for an NRT so that about half of the test takers get each one right. If there are test questions on material which is widely taught and widely learned, pupils are likely to do well on those items, and the items are likely to be rejected for use on an NRT because they do not discriminate among test takers.

In other words, a test which is built to give the most consistent ranking of students (an NRT) will likely not give credit for those aspects of teaching and learning that are generally successful. A test that is built to give a representative measure of how much was learned (a CRT) will give rankings of students that are less consistent, but it should show the results of instruction more readily. Thus the same set of test items may yield both criterion-referenced or norm-referenced meaning, but it will do one of those functions better than the other, depending on how the items are chosen.[1]

One other difference between CRTs and NRTs should be noted: a CRT will typically give more subscores than an NRT of equal length. Strictly speaking, every CRT objective for which a separate score is provided is a separate CRT. A test booklet, then, which covers several CRT objectives is really several short tests.

Use of Tests for Purposes of Instruction and Program Evaluation

The differences between CRTs and NRTs in meaning of scores and in test development imply different optimal uses for each. The fact that CRTs are built around specific instructional objectives makes them especially useful to support instruction. A teacher, school, or district can test the objectives in a CRT battery which are relevant to

---

[1]Hambleton, et al., 1978.

the local program and avoid testing irrelevant skills. Instructional planning for groups of students can then be based on the patterns of specific skills and needs indicated in the test scores. Individual students' strengths and needs can also be diagnosed at the level of teachable skills so that individualized assignments can be made. Similarly, CRTs may be used during the school year to see how well students are progressing in the skills of the local curriculum, so that students may be advanced or helped as needed. In short, the results of CRTs can be directly related to teaching and learning activities and are thus a resource for planning and managing instruction.

The potential for relevance in CRTs has an important effect. Students' scores on CRTs are more likely to reflect the positive achievements which do take place in class than are the scores of a broad survey test which is designed to relate loosely to many varied curricula. This quality of "sensitivity to instruction" is especially timely in an age of educational accountability, since it is important to show as much as possible of the real learning which occurs. In fact, giving teachers and pupils credit for their accomplishments can be seen as a heretofore very underrated purpose of testing.

NRTs, which are designed to differentiate individuals, are most effective for selecting a limited number of very high (or very low) scoring individuals out of a larger pool of available students. They are also capable of giving the most reliable comparison of scores with a national norm. The individual test user will have to decide on the relative importance of these uses of tests--instructional support, giving credit, selection, and comparison with the nation. Guidelines for choosing tests to meet specific needs are given in more detail in Chapter 5.

The use of tests, whether NRT or CRT, in evaluation needs to be placed in context. To many educators, evaluation equals testing.[2] In practice, good evaluation involves a wide variety of management and research techniques aimed at studying the effort, impact, and efficiency of programs at the stages of their preparation, start-up, and ongoing conduct.[3] A major purpose of evaluation is to provide decision makers with information they need to make social programs work well.

In this context, testing is only one part of educational evaluation. Testing can be used at the start of a program as one of several methods for determining curricular needs. During a program, testing can be used as one of several methods for monitoring students' progress in learning so as to help maintain program strengths and modify weaknesses. After a program has been in operation for a reasonable length of time, testing can be used as one of several methods for determining the longer range achievement of the students.

Even with these multiple uses of testing for evaluation, three reservations should be noted. First, it is clear that much inappropriate testing has been done in the name of evaluation.[4] Both the ease of gathering test data and the demand for an accounting of program funds have encouraged an exaggerated reliance

---

[2]Lyon, et al., 1979.

[3]Tripodi, et al., 1978.

[4]Baker, 1978.

on test scores. Second, neither high nor low test scores in themselves are sufficient evidence of program effectiveness. Effectiveness can be judged only in the context of a program's goals and implementation. It would be wrong, for example, to say that an instructional practice or curriculum was ineffective on the basis of low test scores unless it was shown also that the practice or curriculum was adequately put into operation.

The third reservation has to do with the meaning of test results: test scores are not as pure and meaningful as they seem. For a given set of test results, their apparent meaning depends on how they are reported. This point is true both for CRTs[5] and NRTs.[6]

Keeping a perspective on the place of testing in program evaluation, one can more sensibly approach the issue of CRTs vs. NRTs. For credibility in the eyes of whoever commissions an evaluation, a test often needs to have been well validated. For instructional usefulness, a test needs to have close relevance to the program curriculum. At present, standardized tests are generally better validated by field trials than CRTs. But since NRTs are meant to survey a variety of programs, their curricular relevance varies. Also, the methods of selecting questions for NRTs make the items unrepresentative of many curricula. The test user may thus be in the position of choosing the lesser of evils in deciding between tests with strong field test data and tests with curricular relevance. The following section argues that curricular relevance should not be sacrificed when choosing tests.

[5]Barta, et al., 1976.

[6]Linn and Slinde, 1977.

The Importance of Curricular Relevance in Tests

Whether tests are being sought to support instruction directly or to support program evaluation, the single most important feature to consider is the degree to which the objectives of a test match the test user's curriculum. A test may have high reliability, good norms, and other technical virtues; but if the objectives which it tests are not a fair sample of what is being taught, then the test is not a valid measure of that curriculum. Diagnostic tests, for example, give usable information only if the skills on the test are the ones to be covered by the local program. In program evaluation, it is hard to demonstrate the effects of a program by pupils' scores on a test which includes many skills that the program does not attempt to teach. Tests of skills not taught by the local program are at best measures of transfer and at worst measures of I.Q. or general cultural advantage. Low scores on such tests may reveal more about the inappropriateness of the measure than about students' real learning.

Several recent studies show the hazards of using a test that is not closely related to the local curriculum. One study[7] demonstrated that the content of certain standardized tests is not very standard. The authors found that a sample of NRTs of reading achievement reflect the vocabulary of different basal reading series unequally. That is, a given NRT will give better scores for knowing the vocabulary of one reading series than for knowing the vocabulary of others. For the seven reading series examined in the study, the grade level equivalent

[7]Jenkins and Pany, 1976.

score that could be earned by knowing the series' specific vocabulary frequently varied by more than one whole grade depending solely on which test was used, a finding that the authors refer to as "curricular bias in tests."

A second study dealt with reading comprehension.[8] The authors compared the coverage of sixteen separate comprehension skills by three basal reading series and by two widely used norm-referenced tests. In one reading series the proportion of exercise on literal and inferential comprehension was 83% and 17%, respectively, but for the other two series it was about 42% and 58%. Two types of comprehension skills--cloze sentences and words in context--were covered in one or more reading series, but were not included in either test. The cloze sentence exercises represented 24% of the comprehension skills in one reading series, 51% in the second series, and 28% in the third. The words-in-context represented 1% of the comprehension exercises, 1%, and 36%, respectively. Thus the tests failed to credit important parts of these reading programs; and the oversight was unequal across programs.

In a third study,[9] the authors found that four widely used standardized tests of fourth grade mathematics differed markedly from one another in their modes of presenting information and in the nature of the numerical materials used. For example, the proportion of test items using graphs, tables, or figures varied from 15% on one test to 43% on another. The proportion of items using integers varied from 39% to 66% across tests.

In these studies, rather specific skills or aspects of test content were compared. A fourth, more comprehensive study[10] compared tests' coverage of broad objectives for the entire reading and math domains. For this analysis the reading domain was divided into nine non-overlapping objectives and the mathematical domain into thirteen. Coverage of the reading objectives by eight popular NRT series and of the math objectives by seven of the same series was reported for each grade from 1 to 12. The overall trend in the mass of data was that tests differ consistently and widely in the extent to which they emphasize, or even include, the rather general objectives in the two domains.

For the purposes of this discussion, the relevant result of the above mentioned studies is the extent to which the percentage of items per test that are devoted to a given skill actually varies from test to test. The median range in these percentages was 42% for the three most commonly tested reading skills (namely, recognizing meanings of words, literal comprehension, and interpretative comprehension). That is, the test that had the greatest percentage of its items devoted to any one of those skills typically had 42% more of its items measuring that skill than did the test with the smallest percentage of its items devoted to that skill. For the math domain the variation was not as extreme, but still the percentage of items within a test which measured a given objective differed by at least 10% from test to test in 68 out of a possible 156 cases.

The four studies cited were based on an analysis of materials only, not

[8]Armbruster, et al., 1977.

[9]Floden, et al., 1979.

[10]Hoepfner, 1978.

of students' performance on tests. One further study[11] on the effectiveness of traditional and innovative curricula looked at the effects of test content bias on actual test scores. A secondary analysis of more than 20 published research reports led the authors to the conclusion that:

> What these studies show, apparently, is *not* that the two curricula are uniformly superior to the old ones, though this may be true, but rather that *different curricula are associated with different patterns of achievement.* Furthermore, these different patterns of achievement seem generally to follow patterns apparent in the curricula. Students using each curriculum do better than their fellow students on tests which include items not covered at all in the other curriculum or given less emphasis there. (p. 97)

The first four studies show that the content of standardized tests differs and that such tests differ in their correspondence with any given curriculum. The conclusion that such variation in test content could bias the outcome of evaluations, irrespective of students' actual achievement, is confirmed by the last study cited. Thus, if students' scores are affected not only by their actual achievement but also by the mere choice of test, it is essential for tests to be selected so as to maximize their relevance to the local curriculum.

Since curricula differ and since the objectives of ready-made CRTs are not all the same, curricular relevance may be as much a problem for CRTs as for NRTs. In contrast with NRTs, however, CRTs give a separate score for each objective, thus making it easier to distinguish students' performance on program-relevant and program-irrelevant objectives. In some cases, scores on program-irrelevant test items may even be used as a baseline or control measure with which to compare students' achievement on skills that were actually taught.[12]

Issues in Criterion-Referenced Testing...

In the area of CRT, there are many issues on which there is not a consensus. A few of these issues are included here to point out places where the test user may have to make some hard choices. More importantly, this selection of issues is meant to ward off premature complacency about CRT. Just as there are many basic disagreements about standardized testing, much remains to be discovered or decided about criterion-referenced testing.

...Practical issues

Three of these issues are quite practical. First, how shall minimum levels of acceptable performance be set? Ultimately, the choice of a cutting score is determined by the choosers' values; hence it is arbitrary. But the issue remains as to how the arbitrary nature of this process can be made more rational and more politically acceptable. Some methods for setting cut scores are described in the how-to-do-it volume by Hambleton and Eignor.[13]

---

[11] Walker and Schaffarzik, 1974.

[12] Walker, 1978.

[13] Hambleton and Eignor, 1979.

Second, how can test scores be *reported* in a way that is both meaningful and efficient for a CRT that has many separate objectives? For the individual student, test results may exist for 20 or 30 objectives in each of several subjects. Likewise, in program evaluation a large number of objectives and grades may be studied. The problem in both cases involves combining data into a usable, summary form while still conveying significant information. Barta, Ahn, and Castright discuss several methods for dealing with this problem.[14]

Finally, how shall teachers use test scores to make decisions about students? Will tests be a supplement to teachers' judgments about the students or a central tool for decision making? On the one hand, a teacher knows far more about a student than any test can measure. In such cases a test may reveal only what the teacher already knows. Is the test, then, a valuable confirmation of teacher judgment or a costly redundancy? On the other hand, students may have unsuspected needs or strengths that the results of a good test may bring to light. Also, at any given point in the course of instruction, a teacher may need to know which students have reached a pre-set mastery level. In these cases, tests may have a major influence on instructional decisions. Just how to combine test and non-test sources of information to inform the decision making process is a persistent, practical issue for teachers and for people who want teachers to use test scores.

...Theoretical issues

Since this volume is meant to be practical rather than theoretical, these issues will merely be mentioned. The first is whether CRT scores have *construct* meaning or only a *work sample* meaning. In the former case, a CRT score is viewed as measuring an attribute or mental process of the test taker. The different items need to measure the same thing in this case. In the latter case, different items on a test may measure different task components.

A second issue is whether criterion-referenced testing can be applied meaningfully only to achievement or whether CRTs can effectively measure students' attitudes as well. Many writers equate criterion-referenced testing with mastery testing, which excludes measurement of attitudes.

A third issue deals with the importance of field test data for validating CRTs. Some experts argue that a CRT needs only to have a representative sample of items from a well defined domain of behavior in order to be valid. Others hold that field trials are needed for CRTs in order to establish the traditional types of validity. For any CRT that purports to measure psychological traits or processes, including attitudes, validation by field test would obviously be essential. In Chapter 3, this issue receives further attention in the discussion of test characteristics that should be evaluated when selecting tests.

[14]Barta, et al., 1976.

CHAPTER 3
Introduction to the Test Reviews

This chapter introduces the form and content of the test reviews that comprise Chapter 4. First the descriptive component of the reviews is explained. Next the system for the evaluative component is outlined in the form of 21 questions to ask when judging CRTs. Each of the 21 test features and its levels of quality are then explained in detail.

Each test review in Chapter 4 consists of two sections--a description of the test and an evaluation of 21 of its features. The assignment of test characteristics to the descriptive or evaluative category is based on the following rationale: Test features which are likely to affect the test's merit uniformly for most test users are assigned to the evaluative category. Test features which are likely to have very different importance for different test users--cost of the test or format of test items, for example--are assigned to the descriptive category. The intended use of a testing system for such purposes as diagnosis, progress monitoring, program evaluation, and the like, is also described rather than evaluated. Descriptive characteristics affect a test's suitability for the individual user, but such information needs to be evaluated by each user according to local needs and resources.

THE DESCRIPTIVE SECTION OF THE TEST REVIEWS

The descriptive section of each review mentions the intended grades, number of levels, content, intended use, number of objectives, and number of items per objective. The availability of alternate forms is reported here. For any test where pupils *do not* respond on paper, that fact is noted. When the publisher offers supporting materials in addition to the basic test booklet, such as diagnostic and prescriptive aids, these materials are mentioned.

In the descriptive section, the word *levels* refers to levels of difficulty for which separate test forms are provided. Two testing systems may be designed for grades 2 through 7, one having separate test booklets for three broad levels and the other for six narrower ones. Test content is described in terms of broad subject labels such as *reading: word*

attack, or *math: geometry*. Where the publishers have provided such labels, we have used theirs, modifying them only as needed for general familiarity. The reader may locate tests by subject headings in Index B.

Price information is reported in per-pupil terms for tests, answer sheets, and any other major components, for the smallest quantity in which they are available. Note that prices may decline as the size of purchase goes up and that prices change fairly often. The date of the price information is given, but the currency of the costs should be checked before making a purchase choice. Publishers readily provide current catalogs and ordering information. Addresses are given in Index C for the publishers whose tests are reviewed in Chapter 4.

Field test data, if given by the publisher, are described next. The size and composition of pupil populations tested and type of data reported are noted. Details of test administration, such as estimated testing time, special equipment needed, and the need for trained administrators are reported where relevant.

Descriptive information on scoring is given in terms of costs and types of scoring offered. Price information here is also quite changeable. A descriptive category called *Comments* is included for any additional information which does not readily fit in the other categories.

THE EVALUATIVE SECTION OF THE TEST REVIEWS

Each test is evaluated according to 21 dimensions or test features. These 21 features, summarized in question form in Table 1, pages 14-15, fall into three categories:

- Measurement properties (features determining whether the test was constructed according to sound principles of educational measurement).
- Appropriateness for the intended examinees (features determining the suitability of the test for the intended students).
- Usability (features determining the ease with which the test can be administered, scored, and interpreted).

A fourth and critical category--relevance to the test user's curriculum--is not treated here, since the determination of such relevance can only be done with a detailed description of a specific curriculum in hand. Chapter 6 gives assistance in attending to this fourth area of concern.

In reviewing tests, one might compare them on a very large number of features. A recent national sample of school district staff specialists in curriculum, counseling, and testing rated 20 different test characteristics to be *very important* or *crucial* in picking tests.[1] Even a set of 39 characteristics used earlier by CSE[2] is far from complete. A variety of systems for rating tests are used by the books of test reviews listed in Appendix B. Also, a number of other

[1]Dotseth, et al., 1978.

[2]Hoepfner, et al., 1976.

authors[3] have developed guidelines for comparing tests systematically. Since many test features are treated descriptively in this volume, the CSE system for evaluating CRTs looks at only the 21 test features presented in Table 1.

Each question in Table 1 is accompanied by a brief summary of the levels of quality (or standards) which comprise the ratings. Either two or three levels of merit on each feature were used, depending in part on how many different degrees of quality were discernible. Levels were also chosen to try to discriminate among tests, even though this practice resulted in setting the cutting point for a maximum rating at a low level of quality for a few features. The reader should not infer that CSE advocates low standards in tests, but rather should understand that the standards were chosen to try to differentiate tests.

Why should a test user even consider using a test which does not consistently meet high standards of technical merit? Because one other characteristic--relevance to the local curriculum--is more important. Ideally a test buyer would be able to choose from a pool of tests one that is technically sound as well as closely related to the objectives of the local program. When this is not possible, curricular relevance is the less expendable of those two qualities. In this vein, Cronbach[4] has said that precision in test scores is useless if the skills measured by the test are not relevant to the intended decisions.

Ratings of test features in Chapter 4 are expressed in terms of letter grades. Letters are used instead of numbers to encourage test users to weigh the features according to the users' own needs rather than to add the ratings mechanically. Methods for weighing and combining such ratings for the purpose of comparing tests are described in Chapter 5. The letters A, B, and C are used, with A and C being assigned for test features that are divided into only two levels of merit.

In the remainder of Chapter 3, the importance of each of the 21 test features is explained, and levels of merit (standards) are described in greater detail. Casual readers may attend to Table 1 and skip this more technical and detailed explanation of the evaluative criteria. Readers who are involved in selecting tests will profit from the detailed presentation.

NOTE: The information in Table 1 is provided on the inside back cover of this handbook for the convenience of the reader who wishes to refer to it while examining a test review.

---

[3]Cronbach (1970: 186-192), Katz (1973), Popham (1978, Chapter 8), and Hambleton and Eignor (1978).

[4]Cronbach, 1970: 152.

TABLE 1
Key to the Evaluative Sections of CSE Test Reviews*

MEASUREMENT PROPERTIES: CONCEPTUAL VALIDITY

1. Domain Descriptions. How good (i.e., thorough and comprehensive) are the descriptions of the objectives or domains to be tested?
   A. Very good (objectives are thoroughly described)
   B. Adequate (objectives are stated behaviorally but not in detail)
   C. Poor (objectives are loosely described and subject to various interpretations)

2. Agreement. How well do the test items match their objectives?
   A. The match is confirmed by sound evidence
   C. Data are not provided or are not persuasive

3. Representativeness. How adequately do the items sample their objectives?
   A. Items are representative of domains
   C. Item selection is either unrepresentative or unreported

MEASUREMENT PROPERTIES: FIELD TEST VALIDITY

4. Sensitivity. Does conventional instruction lead to test-score gains?
   A. Test scores reflect instruction
   C. Data are not provided or are not persuasive

5. Item Uniformity. How similar are the scores on the different items for an objective?
   A. Some evidence of item uniformity is provided
   C. No data are provided

6. Divergent Validity. Are the scores for each objective relatively uninfluenced by other skills?
   A. Independence of skills is confirmed
   C. Data are not provided or are not persuasive

7. Lack of Bias. Are test scores unfairly affected by social group factors?
   A. Persuasive evidence of lack of bias is offered for at least two groups (e.g., women, specific ethnic groups)
   C. Data are not provided or are not persuasive

8. Consistency of Scores. Are scores on individual objectives consistent over time or over parallel test forms?
   A. Consistency of scores for objectives is shown over parallel forms or repeated testing
   C. Data are not provided

APPROPRIATENESS AND USABILITY

9. Clarity of Instructions. How clear and complete are the instructions to students?
   A. Instructions are clear, complete, and include sample items
   B. Either instructions or sample items are lacking
   C. Both are lacking

10. Item Review. Does the publisher report that items were either logically reviewed or field tested for quality?
    A. Yes
    C. No

*This system for evaluating CRTs is explained in detail in the text. For test features where only two levels of quality are distinguished, the letters A and C are used to indicate the levels.

TABLE 1 (continued)

11. Visible Characteristics. Is the layout and print easily readable?
    A. Print and layout are readable for more than 90% of objectives
    C. At least 10% of objectives have problems in readability

12. Ease of Responding. Is the format for recording answers appropriate for the intended students?
    A. Responding is easy for more than 90% of the objectives
    C. Lack of clarity, crowding, etc., make responding difficult in at least 10% of objectives

13. Informativeness. Does the test buyer have adequate information about the test before buying it?
    A. Yes
    C. No

14. Curriculum Cross-Referencing. Are the test objectives indexed to at least two series of relevant teaching materials?
    A. Yes
    C. No

15. Flexibility. Are many of the objectives tested at more than one level, and are single objectives easy to test separately?
    A. Objectives are varied, carry over across test levels, and are easy to test separately
    B. One feature is missing from variety, carry over, or separability
    C. Two or three of the features are missing

16. Alternate Forms. Are parallel forms available for each test?
    A. Yes
    C. No

17. Test Administration. Are the directions to the examiner clear, complete, and easy to use?
    A. Directions are clear, complete, and easy to use
    C. One or more of the above features are missing

18. Scoring. Are both machine scoring and easy hand scoring available?
    A. Yes
    B. Easy, objective hand scoring is available, but no machine scoring
    C. Hand scoring is not easy or objective; or only machine scoring is offered

19. Record Keeping. Does the publisher provide record forms that are keyed to test objectives and are easy to use?
    A. Yes
    C. They are not included or not keyed to test objectives

20. Decision Rules. Are well justified, easy-to-use rules given for making instructional decisions on the basis of test results?
    A. Yes
    C. Decision rules either are not given, not easy to use, or not justified

21. Comparative Data. Are scores of a representative reference group of students given for comparing with scores of pupils in the test user's program?
    A. National norms, criterion group data, or item difficulty values are provided
    C. These are not provided or are not clearly representative

MEASUREMENT PROPERTIES: CONCEPTUAL VALIDITY

A test score is not an end in itself; it is a sign or indicator of something more important. A score may give a prediction about the pupil's future performance, or it may give an estimate of how the test taker is likely to perform on a larger set of possible items from which the test items are drawn. In the latter case, that pool of possible test items is called the *domain* (or *criterion*). A pupil's score on a CRT thus gives an estimate of how the pupil is likely to perform with respect to the population of all such items.

One essential step in making the scores of a CRT meaningful is to describe the criterion pool of items clearly. First, a clear description enables teachers to teach the skill or attitude that is described. By providing a practical target for instruction, the description makes the score on such a test userful for diagnosis and prescription. Second, a clear test description can help to demystify testing by telling consumers of test results just what was tested. The description thus gives meaning to the score. In most of the tests covered in this book, the descriptions of the criterion behaviors take the form of instructional objectives. When "a test" is referred to, it means a group of items that provides a separate score. One CRT test form may thus contain several tests.

Since the items of a CRT are supposed to test the skill or attitude as set forth in the description of the criterion, the validity of a CRT depends on the extent to which the items actually fit the test description. This type of validity is often referred to as *content validity*, but that term is too narrow. Popham[5] has suggested the phrase *descriptive validity* so as to apply not only to CRTs in the cognitive domain but also to those in the psychomotor and affective domains, where process or action may be more relevant than content. A description of the criterion that clearly specifies what should and should not be included is an essential link in determining whether a CRT has this type of validity.

1. Domain (or Criterion) Descriptions

Background

For test buyers, thorough domain descriptions have practical potential. A test user could compare the curricular relevance of two or more CRT batteries by seeing how well their domain descriptions match the local program, rather than by having to examine the test items directly. A full CRT description will consist of a set of instructions to the test writer that prescribes the content, format, and mode of responding for all of the possible test items. Directions for making up multiple choice options, for scoring free responses, and for sampling items from the criterion item pool will also be given. Much of this information goes beyond subject matter content.

It is obvious that detailed domain descriptions are technical documents, too lengthy and detailed in their entirety to be efficient either for planning instruction or for reporting grades. But the detailed descriptions can include brief statements for teachers and parents in a form like behavioral objectives.

---

[5]Popham, 1978.

Levels of Quality

Level A.[6] Content, format, response mode, and sampling rules are described thoroughly enough so that (a) different test writers should produce equivalent tests by following the description, or (b) for any test item or set of items, it is clear whether they fall inside or outside the intended domain. The names of three types of test description that are most adequate are *item forms*, *amplified objectives*, and *domain specifications*.

Level B. Content, format, and response mode are described, as in a behavioral objective. Rules for sampling items are not given, or there is so much slack in the limits of content, format, or response mode that differing tasks could still fit the description. Tests based on such descriptions are objectives-based.

Level C. The test is described in terms that give little indication of the content, format, and response mode of the test items. General skill category labels, such as *reading comprehension*, *word attack skills*, or *basic mathematical operations*, are at this vague level of description. Many different types of test items will fit a description as general as this one. Since these descriptions give little indication of what the criterion behaviors are, tests with such descriptions are scarcely criterion referenced.

[6]Appendix D has an example of a domain description which would receive a level "A" rating.

2. Agreement of Items with their Test Descriptions

Background

The *domain descriptions* of feature #1 above are a test maker's intentions for constructing tests. It is still necessary to show that the intentions were carried out. Features #2 and #3 deal with this issue. Feature #2 asks whether the test items are accurately described by the test description. If they are not, then the items test something else and the test is invalid. Technical terms that are used to refer to the concept of agreement include *item-objective congruence*, *content validity*, and *descriptive validity*.

Levels of Quality

Level A. Sound evidence of agreement is offered and described in enough detail to evaluate. The test developer gives a detailed account of either how the items were generated from the description of the criterion behaviors or how qualified judges confirmed the fit of the individual items to the description.

Level C. No evidence of agreement is offered; or evidence is mentioned but not described in enough detail to evaluate; or evidence is described in detail but is flawed.

3. Representativeness of the Items

Background

Rarely is a test score of interest for its own sake. Test scores are used as observable indicators of more important things that are difficult or impossible to observe directly. For example, students'

scores on any achievement test are used to indicate their mastery of a total set of possible questions on the subject matter. It is rarely possible to test the total set. Likewise, a person's performance on a test of intelligence or personality is used as an indicator of how the person will act in more natural situations. For a test score to be an accurate indicator, the test items must not be chosen in a biased manner. In other words, the items must be chosen in a way that allows for generalization from the test score to the intended total set of behaviors. If the selection process is biased, unplanned, or unrepresentative, then the total set of behaviors that the test score represents cannot be determined.

Levels of Quality

Level A. The test developer reports that the items were selected either randomly from the set of questions possible under this objective or, if there are components in the domain, by stratified random sampling.

Level C. No account is given of how the test questions were chosen from the set of questions possible under this objective; or items were selected in a biased or unrepresentative fashion. Items are not representative if the item selection process systematically excluded those that failed to discriminate high and low scoring individuals in a group of students who have a common instructional background.

MEASUREMENT PROPERTIES:
FIELD TEST VALIDITY

Authorities in the field of CRT agree that conceptual validity is necessary for a good criterion-referenced test. They do not agree, however, on the necessity for empirical (data-based) validation of CRTs. CSE takes the position that the two types of validity are interdependent; both are necessary for confirming that a test measures what it claims to. Without validation by field trials, a test that appears to be conceptually sound may give measures that are not consistent (test feature #8), that do not reflect the relevant learning (#4), that are of an unintended mixture of behaviors (#5), that are affected by skills or attitudes other than the intended one (#6), and that are biased (#7). Without meeting the standards for Conceptual Validity, on the other hand, a test may be an unrepresentative measure (#3) of the wrong criterion (#2) or of no identified criterion at all (#1).

4. Sensitivity to Learning

Background

Students' scores on a test may or may not reflect their actual learning of the skills which the test purports to measure. To the extent that the scores do, the test is said to be sensitive to learning. This feature for judging the merits of tests is not universally accepted, in part because it is usually called sensitivity to *instruction*. The objection is that a test may not show any effects of instruction because the given instruction did not have any effect. Thus, when a small sample of students in a field test does not appear on a posttest to have

benefited from instruction, that result is not necessarily the fault of the test.

The objection is well taken as far as it goes. However, consumers of tests need to know that the test does reflect positive effects of instruction in a fair proportion of classrooms. If it does not, either the test is insensitive or the test content is not teachable by current methods. In either case, such a test will not be useful.

Demonstration of sensitivity to learning under one form of instruction or with one type of pupil will not guarantee sensitivity to all forms of instruction or for all types of pupils. The test developer should describe the type(s) of instruction and pupil used in the field tests so that test buyers can decide if the test is likely to be sensitive in their own setting.

There are serious technical problems in measuring change, and there is not yet a consensus on how to prove a test's sensitivity. This test feature was evaluated here simply by asking: Does the test developer offer *any* evidence of a test's sensitivity that is free from the well established problems in measurement (e.g., unreliability, the effect of experiences outside the school)? Such data must be provided for each separately scored skill.

Levels of Quality

Level A. The test has been found to reflect learning in a representative sample of students following an ordinary (in terms of time, intensity, and resources usually available for the particular subject) course of instruction. The course of instruction is clearly aimed at the criterion behaviors. The well established problems in measurement are not present in the study.

Level C. No information is given on the sensitivity of the test to student gains; or evidence suggests that the test suffers from well established problems in measurement[7]; or the gains cited are not statistically dependable; or the successful teaching method was not described.

5. Item Uniformity

Background

This feature deals with whether a test (i.e., each separately scored set of items) measures a uniform, coherent skill or attitude. If the test does not, then it measures a mixture of things. A CRT that is a uniform measure is a better test, with the following exception. In some cases, the definition of the criterion behaviors identifies different components or levels of difficulty. For example, a phonics test might deal with consonants, the differing categories of consonants (e.g., stops, liquids, nasals, fricatives) being identified as components of that phonic skill. Such a test should show uniformity within each category, but not necessarily within the whole test of several categories. When such a test measures a mixture of things, the mix is planned. An accidental lack of uniformity results when the items unintentionally call for different skills or attitudes. It is a sign that the description of the criterion is defective, for the test does not measure what it purports to measure.

Uniformity or coherence of a CRT is shown by measures of the extent to which all items for a given skill function alike. The more a student's score on one test item is

[7]Campbell and Stanley, 1963.

similar to his scores on the other items, the more uniformity the test has. In classical test-score theory, *factor analysis*, *inter-item correlations*, and *part-whole correlations* give measures of uniformity.

Levels of Quality

Level A. At this early stage in the development of CRTs, any numerical evidence of item uniformity will be accepted if it is reported for groups of items testing single objectives. The data must be based on students' responses to the test items.

Level C. No numerical evidence of item uniformity is given at the level of the individual objective, or only judgmental evidence is given.

6. Divergent Validity

Background

This feature deals with whether the scores on a test are relatively uninfluenced by achievements or attitudes that the test is not supposed to be measuring.[8] If the scores on the test are relatively uninfluenced by other, unintended factors, then it is a test of something distinct and has divergent validity. For example, the more that scores on a test of reading comprehension are influenced by general knowledge, apart from the examinees' understanding of the test, the less the divergent validity of the test. For a math test to have divergent validity, its language must be simple enough so that pupil errors are not reading errors.

Divergent validity, or separateness, can be confirmed by traditional methods--factor analysis, correlation studies among measures of separate behaviors--or by experimental evidence that scores on a test respond to a relevant treatment while scores on certain other tests do not.

Levels of Quality

Level A. Evidence of divergent validity is given showing the CRT's scores to be independent of scores on tests of other supposedly unrelated achievements or attitudes.

Level C. No evidence of divergence is offered, or the evidence is not detailed enough to judge. There is evidence of contamination (for example, high correlations of CRT scores with other I.Q. scores or scores of verbal aptitude).

7. Lack of Bias

Background

This feature is concerned with how different groups of students--for example, different ethnic groups--perform on a test. It does not deal with the surface content of test questions. Bias has been common enough in testing so that it is unwise to assume that it is absent from current tests. Hence a demonstration of lack of bias is required to confirm a test's validity for major social groups.

A test is biased for a given group of students if it does not permit them to demonstrate their skills or attitudes as completely as it permits other groups to do so. Such a test is invalid for that group. The subject of bias is surrounded with controversy, in part because social injustice for large numbers of students can result from biased tests.

[8] Campbell and Fiske, 1959.

Levels of Quality

Level A. Evidence of lack of bias is offered for at least two of the following groups: women, Blacks, and students from Spanish speaking backgrounds. Lack of sizable item-by-group interactions is one form of evidence. A second is similarity across groups of the other data for empirical validity (features #4-8).

Level C. No evidence of lack of bias is offered, or evidence is offered but not persuasive. A difference in the average scores of ethnic or other groups by itself will not be considered evidence of bias.

8. Consistency of Scores

Background

A test is *consistent* if the difference in a student's scores on two occasions is due to a real change in achievement or, for affective measures, in attitude. If a student's scores change as a result of the vagueness of the instructions, variations in testing conditions, or other factors aside from real learning, then the test's scores are not consistent. Changes in scores due to irrelevant factors make the scores of any one occasion suspect. The more that a test's scores reflect real learning, and not irrelevant factors, the more *consistent* it is.

Consistency measures used with norm-referenced tests include estimates of *test-retest reliability* and *alternate form reliability*. The traditional reliability estimates often are not suitable for CRTs, and thus the use of the broader term consistency. When CRTs are to be used in a pass/fail fashion, consistency should be shown for the pass/fail judgments.

Consistency data are necessary to show that a test's scores are dependable, but not many such studies have yet been done on CRTs. In principle, consistency may vary over a wide range; but current CRTs differ more on whether they report consistency data at all than on the values reported. At this point, the reporting of any such data is seen as a positive step in test development and a step toward truth in packaging.

Levels of Quality

Level A. Data are reported on the consistency of students' scores. Either consistency of individuals' scores over repeated testing or consistency of individuals' scores on different forms of the test will be credited.

Level C. No consistency data are given.

APPROPRIATENESS AND USABILITY

The effects of features #9-12 would show up in the validity and consistency data for a test. Because little information is yet available on the measurement properties of CRTs, and because features #9-12 may cause problems in giving a test, they are treated separately here.

9. Clarity of Instructions to Students

Background

The instructions to students must describe all aspects of the task in language that is suited to the intended age or grade levels. Sample items that are both typical and clear should be given both for practice and clarification.

Levels of Quality

Level A. The instructions to the intended test takers are clear and complete, and a sample item is provided.

Level B. Either the instructions are not appropriate or the sample items are lacking.

Level C. The language of the instructions is too advanced or otherwise inappropriate for the intended group; or instructions are incomplete or hard to follow; or a sample item is not given.

10. Item Review

Background

Test items are appropriate if they are understandable, have at least one correct answer, give credit for all correct answers, do not give away the correct answer, and are otherwise free from technical flaws. Two kinds of evidence are considered here--namely, test developers' reports that the items were logically (i.e., judgmentally) reviewed, or that they were reviewed through field testing.

Levels of Quality

Level A. The test developer reports that item quality was reviewed independently of item writing.

Level C. The test developer offers no evidence that item quality was checked apart from the process of original item generation.

11. Visible Characteristics of Test Materials

Background

The visible characteristics of test materials should make it easy for students at the intended levels to use the materials. Tests were examined for the details of layout, organization, and clarity mentioned under Level C below.

Levels of Quality

Level A. More than 90% of the objectives are free of the flaws listed under Level C.

Level C. At least 10% of the objectives have one or more of these flaws: print of pictures is unclear, items are too close together, stems and responses are not clearly grouped, sequence of items is easy to lose, there is little blank space for math work, the page is cluttered, item numbers are not easy to pick out, information needed to answer a question is unnecessarily spread out.

12. Ease of Responding

Background

A test should be formatted so that students' scores are not affected by difficulties in recording their answers. Answer sheets or other spaces for responding are judged for not only the attributes mentioned in feature #11, but also for the amount of space provided for answers.

Levels of Quality

Level A. More than 90% of the objectives are free of the response material flaws described under Level C.

Level C. Answer sheets or other response sheets for at least 10% of the objectives have one or more of these flaws: print is unclear, items are too close together, item numbers are not easy to pick out, answer spaces are too small.

13. Informativeness of Materials for the Prospective Buyers

Background

Some publishers make it easier for a prospective buyer to decide whether or not to purchase a test by providing complete, easy-to-use information on the materials. There are usually two stages in test purchase: ordering sample materials and ordering the testing package itself. Since the presence and quality of technical information on test development is covered in features #2-8, it will not be counted here.

The issue here is whether the prospective buyer knows what the testing package will consist of before investing in it. This feature is more important in weighing the more costly CRT systems, where the prospective user will be less willing to buy the system without first having an opportunity to examine it.

Levels of Quality

Level A. Either the whole system is available on approval, or the following possibilities are available to the prospective purchaser as part of the publisher's promotional effort: specimen sets of sample pages and instruction can be obtained; test copies can be purchased in any quantity; a complete listing of the test's objectives is provided before purchase; information on ordering of original and replacement materials is clear and complete; replacement materials may be ordered separately; information on returning unused materials is clear; pertinent information is available without buying the testing package--the instructions to students and test users, the physical characteristics of the test, instructions and materials for recording answers, the number of separate test forms available, decision standards and comparative data, time required for testing, the training needed to give or interpret the test.

Level C. Promotional materials do not give enough information for deciding whether to order specimen sets in one or more of the respects listed under Level A.

14. Curriculum Cross-Referencing

Background

A testing package is easier to coordinate with local curriculum and instruction if it includes an index relating its objectives to specific teaching materials. Such an index

can be used to guide test selection and to help teachers locate alternative instructional materials. For either purpose the user will have to verify that the indexed instructional materials adequately cover the same skills as the test and the local curriculum.

Levels of Quality

Level A. Indexing of the test's objectives to two or more publishers' teaching materials is provided in detail (e.g., specific units in specific texts).

Level C. No curriculum cross-referencing is provided.

15. Flexibility in Choosing Objectives

Background

A test or testing system is adaptable to a range of local needs--for example, individualization--if it covers a variety of objectives and tests them on separate forms. A testing system which combines a fixed set of objectives does not give the user as much control over testing. Flexibility is also provided by a system which gives tests of the same core objectives at more than one test level. Such a system does not have the same test items on forms that differ only in the level marking, but has tests of the same skills with content and illustrations suited to the different levels.

Note that it is not fair to compare large-scale testing systems with smaller ones that do not try to cover the same range of skills or grades. If the user is looking for a highly specific test, this criterion may not be relevant. Also, the cost of such flexibility will be an important consideration to the test buyer.

Levels of Quality

Level A. All of the following features are present: variety in the objectives, separate forms, or grade level flexibility of core objectives.

Level B. One of the features mentioned under Level A is missing.

Level C. The test or system provides a narrow range of objectives and prints several of them together on the same test form. Core objectives are available in materials appropriate to only one grade level.

16. Alternate Forms

Background

When a testing system has alternate forms, the user can give independent retests to the same students. If retesting is done with the same form that was used for the original test, students' scores are likely to be influenced not only by their learning of the subject matter but also by specific memory of the first testing. This latter influence invalidates the retest scores. With alternate forms, pre- and posttesting or repeated posttesting can be done without this invalidating carryover effect.

Levels of Quality

Level A. Two or more forms with non-overlapping sets of items are available for each test.

Level C. Only one form is available for each test.

17. Test Administration

Background

A test is more practical if the instructions to the examiner are clear, complete, and well organized. With good instructions, the testing is not only easier, but the testing conditions are also more uniform.

Levels of Quality

Level A. Instructions leave little room for misunderstanding by the examiner and are complete and easy to use.

Level C. Instructions to the examiner are hard to find or follow. They are vague, ambiguous, not complete, not all in one place, or not logically ordered. Or, the copy in the manual is unclear.

18. Scoring

Background

A test is more practical if it can be scored easily and objectively and if the test user is not limited to one method of scoring. Hand scoring is easy if scoring templates or other well organized keys are provided.

Levels of Quality

Level A. Both machine and easy, objective hand scoring options are available.

Level B. Only hand scoring is available, but it is objective and easy.

Level C. Hand scoring is difficult, or arbitrary, or requires special training. Or, scoring requires the expense of special machines on site or the delay of sending students' responses out for scoring.

19. Record Keeping

Background

Good records of student performance are an important part of classroom management and of meeting accountability requirements. CRT systems have the potential for making record keeping burdensome because they often have large numbers of objectives. A testing system is more practical when it has forms for recording students' test scores that are easily keyed to the objectives, easy to maintain, and easy to interpret.

Levels of Quality

Level A. Usable forms for record keeping are provided.

Level C. Either teachers must create their own record forms, or the testing system's forms are not easily keyed to the objectives, easy to maintain, or easy to interpret.

20. Decision Rules

Background

Tests may be used to make decisions about students. Tests should be constructed in a way that allows decisions to be made with confidence and ease. The information for decision making should be easy to find, easy to use, and well justified. Although the choice of cutting scores for passing or mastery should be left to the local test user, the publisher should give an indication of the consequences of choosing different cutoffs.

Relative costs and gains will affect the choice of a cutting score. Where a prerequisite skill is being tested, it may be preferable to hold back a few students who have actually mastered it in order to avoid passing ones who have not. In other cases, holding students back may be more costly than advancing students before they attain mastery.

One aspect of test design that affects the decision rules has not been covered previously--namely, number of items per test or per objective. For several reasons it is important to have more than just a few items per objective. First, there must be enough items so that occasional misreading of questions by students will not result in unwarranted failures. Second, there must be enough items so that chance effects, like guessing, do not result in unwarranted passing. Ideally there will be enough test questions so that three levels of attainment can be identified: clear pass, clear fail, and an area of uncertainty. Finally, a sufficient number of items on a test is a protection against misjudging individual students' scores in case students share an occasional answer.

Since statistics for CRTs are still largely under development, including the statistics for decision making, only two general levels of merit are applied here.

Levels of Quality

Level A. Decision rules that are easy to find and use are provided along with a rationale for their use.

Level C. Decision rules are not provided, or they are provided without justification, or they are hard to find or use.

21. Comparative Data

Background

Authorities disagree on whether the intent of criterion-referenced testing is undermined by providing comparative (that is, norm-referenced) interpretations of CRTs. But test scores are not easy to interpret, and the more information that can be provided about them, the easier it is to understand and explain them to others. Thus CRTs that offer both absolute and relative interpretations of scores are seen as more practical than ones that have only the former.

Test users should recall that NRTs are designed to provide stable rankings of students in that they consist of test items that spread out the scores of test takers. A well designed CRT (features #1-3) is likely to provide less stable rankings because items are sampled to be representative of the skills or attitudes, and because the number of items on a test (i.e., on a single objective) is likely to be smaller.[9]

Note that comparative data need not be percentile norms. Average percent correct could be given for various reference groups. Grade level equivalents are not acceptable owing to their many problems.[10]

Levels of Quality

Level A. Acceptable comparative data are based on the responses of at least several hundred students in a nationally representative sample.

---

[9]Hambleton, et al., 1978.

[10]APA, 1974; Tallmadge and Horst, 1976; Linn and Slinde, 1977.

Percentile norms, data for well identified reference groups, or summaries of performance of students in the target grades are suitable. These data are easy to find and interpret in the user's manuals.

Level C. Comparative data are either not provided, or consist only of grade level equivalents, or are based on the responses of a small or unrepresentative sample of students. Or these data are hard to find and interpret.

CHAPTER 4
CSE Criterion-Referenced Test Reviews

A key summarizing the rating system used in the following test evaluations can be found on the inside back cover of this handbook for the convenience of the reader who wishes to refer to it while examining a test review.

ANALYSIS OF SKILLS (ASK) - Language Arts

Scholastic Testing Service, 1975

DESCRIPTION

The ASK-Language Arts is a six-level battery of tests designed for diagnosis and prescription in grades 2-8. It covers a total of 73 objectives in the following broad areas: capitalization and punctuation, usage, sentence knowledge, and elements of composition. There are 36 to 58 objectives per level, each with three multiple choice items. Items within each objective are "spiraled" so that the easier ones come earlier in the test form, the harder ones in the latter part of the test.

PRICES

Two systems for ordering materials and services are available, a conventional one and a "lease-score" system. For materials that are to be retained, the reusable test booklets are 62¢ each in packages of 20; answer sheets, 17¢ in sets of 50; examiners' manuals are 70¢; and a general manual is $2.00. Individual student record forms which are included in the basic scoring service are 12¢ each in sets of 20 when purchased separately.

Under the "lease-score" system, the purchaser returns all test materials to the publisher and pays only the costs of data processing, reporting, and transportation. Specimen sets are $2.00. Date of prices: 1978.

FIELD TEST DATA

The developer has unpublished data showing that each level of ASK-Language Arts was field tested on a median of 145 pupils. The data are mostly in the form of numbers and percents of pupils picking each response choice. The field test is not described.

ADMINISTRATION

The ASK tests are designed for group administration. Tests are untimed, but the publisher recommends scheduling two sessions for any one level and estimates the total testing time as 1 minute per item, i.e., 108 to 174 minutes for Levels 2-3 to 7-8.

SCORING

The basic leasing service costs $1.10 per pupil and includes three class record sheets, one set of individual student record forms and labels, one interpretative brochure per class, and a content outline "per level per class." The basic scoring cost for those who purchase materials is 78¢ per pupil and includes everything except test booklets and answer sheets.

ANALYSIS OF SKILLS (ASK) - Language Arts

Scholastic Testing Service, 1975

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review. It is not clear from the unpublished field test data that they were used for item revision.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. Core objectives are covered, each at several levels. The intentional "spiraling" of items within objectives makes single objectives hard to test separately.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. Hand scoring is difficult. Machine scoring is offered.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data. When the scoring service is used, local and national norms for the content areas and for the total scores may be printed on the class record sheet by pupil. The composition of the national norm group is not described.

ANALYSIS OF SKILLS (ASK) - Mathematics

Scholastic Testing Service, 1974

DESCRIPTION

ASK-Math is a seven-level battery of diagnostic tests for pupils in grades 1-8 covering these categories of skills: computation, concepts and problem solving, and applications. There are 44 to 58 objectives per level with three multiple choice items per objective. Items within each objective are "spiraled" so that the easier ones come earlier in the test form, the harder ones in the latter part of the test.

PRICES

Two systems for ordering materials and services are available, a conventional purchase system and a "lease-score" system. For materials that are to be retained, reusable test booklets are 62¢ per pupil in packages of 20 and answer sheets 17¢ in sets of 50. The examiner's manuals, one for Level 1-2 and one for Levels 2-8, are 70¢, and a general manual is $2.00. Individual student record forms, which are included in the basic scoring service, are 12¢ each in sets of 20 when purchased separately.

Under the "lease-score" system, the purchaser returns all materials to the publisher and pays only the costs of processing, reporting, and transportation. Specimen sets are $2.00. Date of information: 1978.

FIELD TEST DATA

The developer has unpublished data showing that each level of this battery was normed on a median of over 1100 pupils. The field test is not described.

ADMINISTRATION

The ASK-Math tests are made for group administration. The tests are untimed, but the publisher estimates the total testing time to be 180 minutes for Level 1-2 and 130 minutes for each other level.

SCORING

The publisher does not recommend hand scoring. The lease-scoring service, at $1.43 for Level 1-2 and $1.10 for the other levels, provides individual pupil folders with score labels, an interpretive brochure, and objectives-based and normative scores for individuals and for the group.

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions. Instructions for level 1-2 appear too complicated, but other levels do not.

A (C) 10. Item Review. It is not clear from the unpublished field test data that they were used for item revision.

A (C) 11. Visibility. At level 1-2, there is a problem of crowding and print size. Other levels appear satisfactory.

(A) C 12. Responding.

A (C) 13. Informativeness. Prices and contents are not clearly laid out in the catalog.

A (C) 14. Curriculum Cross-Referencing.

A B (C) 15. Flexibility. Although core objectives are covered at several levels, about one-third of the items are repeated across levels. The intentional "spiraling" or items within objectives makes single objectives hard to test separately.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. The publisher describes hand scoring as difficult. Machine scoring is offered.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data. Normative scores are offered, but the composition of the norm group is not described.

ANALYSIS OF SKILLS (ASK) - Reading

Scholastic Testing Service, 1974

DESCRIPTION

The ASK-Reading is a four-level battery of tests for pupils in grades 1-8 which cover the following major skill areas: word analysis, comprehension, and study skills. There are 43 to 48 objectives per level with three multiple choice items per objective.

PRICES

Two systems for ordering materials and services are available, a conventional purchase system and a "lease-score" system. For materials to be retained, reusable test booklets are 62¢ per pupil in packages of 20 and answer sheets 17¢ in sets of 50. Examiner's manuals are 70¢ and a general manual for reading and math is $2.00. Individual pupil record forms, which are included in the basic scoring service, are 12¢ each in sets of 20 when purchased separately.

Under the "lease-score" system, the purchaser returns all test materials to the publisher and pays only the costs of processing, reporting, and transportation. Specimen sets are $2.00. Date of information: 1978.

FIELD TEST DATA

The developer has unpublished data showing that each level of ASK-Reading was field tested on a median of about 200 pupils. The data are mostly in the form of number and percent of pupils picking each response choice. The field test is not described.

ADMINISTRATION

The ASK-Reading tests are made for group administration. The tests are untimed, but the publisher estimates the total testing time to be between 2 and 2½ hours per level.

SCORING

The publisher does not recommend hand scoring, but answer keys are provided in the manuals of directions for the lowest level of test, Level 1-2. The basic scoring service is $1.43 per pupil for Level 1-2 and $1.10 for the other levels.

ANALYSIS OF SKILLS (ASK) - Reading

Scholastic Testing Service, 1974

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions. Instructions for Level 1-2 appear too complicated, but the other levels do not.

A (C) 10. Item Review. It is not clear from the unpublished field test data that they were used for item revision.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. Overlap of objectives across levels is provided, but all items for a level are on one test form.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. Hand scoring is discouraged. Machine scoring is available.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Decision rules are given, but with little support.

A (C) 21. Comparative Data. When the scoring service is used, local and national norms are provided. Composition of the national norm group is not described.

BASIC ARITHMETIC SKILL EVALUATION (BASE) AND BASE II	Imperial International Learning Corporation, 1972

DIRECTIONS

BASE and BASE II are two diagnostic and prescriptive systems for math in grades 1-6 and 7-8 respectively. BASE has six levels of 16 to 23 objectives each and covers the following skill areas: numeration and operations with whole numbers, fractions, money, measurement, geometry, story problems, decimals, and percents. BASE II measures objectives in operations with integers, fractions, decimals and percents, and story problems. For both batteries, there are three multiple choice items per objective. Reference guides to prescriptive materials are a part of the system. The cards for posttesting individual pupils are a type of alternate form.

PRICES

The BASE system for grades 1-6 sells for $229 and includes for each grade level a cassette tape of instructions, a reference guide, consumable tests for 30 pupils, 30 student profile sheets, and a set of posttest cards. The price per grade level is $39.50. With BASE II, which cost $54.00 separately, the complete system is $269.00. The cost for replacing tests and profile cards for 30 pupils is $19.50 for each primary grade and $21.50 for BASE II. Date of information: 1978.

FIELD TEST DATA

Field testing of BASE II is mentioned but not described.

ADMINISTRATION

The BASE system is designed for group administration or self-administration, both with the aid of tape recorded instructions. Each level of the tests for grades 1-6 is estimated to take 1-1½ hours. Three class periods are suggested for giving BASE II.

SCORING

The carbonized answer sheets are self-scoring.

BASIC ARITHMETIC SKILL EVALUATION (BASE) AND BASE II

Imperial International Learning Corporation, 1972

MEASUREMENT PROPERTIES

A (B) C 1. Description. No data.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

A C 13. Informativeness. The rating here will depend on the contents of the specimen set: Does it include a copy of the objectives?

(A) C 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. There is good carry over of objectives from level to level, but separate testing of objectives is not easy.

A (C) 16. Alternate Forms. The posttest cards do not make group posttesting practical, although they may be considered an alternate form.

(A) C 17. Administration.

A (B) C 18. Scoring. The print on the self-scoring duplicates is often very faint.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

BASIC WORD VOCABULARY TEST
by Dr. Harold J. Dupuy

Dreier Educational Systems

DESCRIPTION

The Basic Word Vocabulary Test is a 123-item multiple choice test designed for test takers from 4th grade through Ph.D. level. Item stems are root words or simple phrases with a root word underlined. These words were selected from a 1% sample of words that are common to four major unabridged dictionaries. Eliminated from the sample were foreign, slang, archaic, and technical words. Response choices are single words or short phrases. Items are arranged in order of increasing difficulty.

PRICES

A package of 40 test booklets, the examiner's manual, and scoring key sells for $4.95. The specimen set, at $2.95, contains an examiner's manual and sample test for this and each of four other tests by Dreier. Date of information: 1978.

FIELD TEST DATA

A developmental pretest was done on 148 people ranging in age from 11 to 61. After revision, the final form was administered to 3,100 students in grades 1 through 12 in the public schools of Fairfax, Virginia. The examiner's manual gives percentiles and grade level equivalents for grades 3 through 12, and percentiles for college and graduate students. These latter scores were derived by extrapolation, since the norming population went only through the 12th grade. IQ-like scores based on vocabulary alone, called the Vocabulary Development Quotient, are also given. Raw scores provide estimates of pupils' mastery of the 12,300 word "basic vocabulary." The detailed technical manual, DHEW Publication No. (HRA)74-1334, is reprinted in ERIC as ED 094 373.

ADMINISTRATION

The BWVT can be used as a group test. Pupils read the test words to themselves and stop where indicated on the test form (e.g., 4th graders after 68 words). Estimated testing time is 20 minutes or less.

SCORING

Tests are scored by hand with an overlay or by machine. The user is invited to write Dreier for information on machine scoring.

BASIC WORD VOCABULARY TEST
by Dr. Harold J. Dupuy

Dreier Educational Systems

MEASUREMENT PROPERTIES

(A) B C 1. Description. The criterion pool of words is described, and rules are given for generating distractors and correct answers.*

A (C) 2. Agreement. Although the technical manual does not report that item-domain agreement was verified independently, the careful domain description makes it likely that agreement is quite high.

(A) C 3. Representativeness.

A (C) 4. Sensitivity.

(A) C 5. Item Uniformity. Correlations of scores on 40-item subsets of the total test ranged from .95 to .97.

A (C) 6. Divergent Validity. The publisher's suggestion to get an IQ-like score indicates that the test is more an IQ test than an achievement test.

A (C) 7. Bias. The technical manual says that tests like this *should* reveal the effects of cultural deprivation so that the need for remediation can be identified.

A (C) 8. Consistency.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review. Items were revised after field testing.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A B C 15. Flexibility. For this one-objective test with a graded vocabulary, flexibility is not relevant.

A (C) 16. Alternate Forms. The technical manual gives three parallel forms which are subsamples of the complete test.

(A) C 17. Administration.

(A) B C 18. Scoring. A transparent overlay or machine scoring may be used.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. The instructional implications of a score on this test are not clear.

A (C) 21. Comparative Data. A well-defined but local sample of about 275 students per grade provided the norms.

*The BWVT includes many very rare words, over 30% of the stem words not appearing in the Thorndike-Lorge word count. Although a criterion-referenced interpretation of this test is possible, the criterion pool of words is not a useful one for general education. A number of the words will be much less familiar in some regions of the U.S. than in others.

BEGINNING ASSESSMENT TEST FOR READING J. B. Lippincott Company, 1975

DESCRIPTION

The tests in this battery measure the development of skills which are related to early reading instruction. The skills covered include vocabulary, visual and auditory discrimination, classification, rhyming, sequencing, riddles, letter recognition, sound-letter correspondences, picture-word and picture-sentence matching, spelling, sentence completion, oral production and comprehension, and color naming. A 41-item placement test measures 12 objectives with 2 to 6 items each. The comprehensive test measures 19 objectives with 6 to 26 items each. Responses are spoken, written, and selected from multiple choices.

PRICES

When purchased in the boxed set for 35 pupils, the cost is $21.78. This set includes consumable practice tests, placement tests, and comprehensive tests, manuals for each of these tests, and record forms. Date of information: 1978.

FIELD TEST DATA

Field testing is mentioned, but results are not reported.

ADMINISTRATION

These tests are designed for administration by a teacher to groups of 8 to 15 children, except for two subtests that require oral responses. Estimated testing time is 30-40 minutes for the placement test and 60 minutes for the comprehensive test.

SCORING

Hand scoring is done with replicas of the pupils' answer pages.

BEGINNING ASSESSMENT TEST FOR READING J. B. Lippincott Company, 1975

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data, but attention was given to avoiding stereotypes in item development.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review. Items were screened on the basis of a small field test.

(A) C 11. Visibility.

(A) C 12. Responding.

A (C) 13. Informativeness. No specimen set.

A (C) 14. Curriculum cross-referencing.

A (B) C 15. Flexibility. Not entirely relevant, since the test is designed for one level. Items for each objective are printed on different pages.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Rules are provided, but without support. Interpretative guidance for prescription or placement is not given.

A (C) 21. Comparative Data.

CARVER-DARBY CHUNKED READING TEST Revrac Publications, 1970

DESCRIPTION

The Carver-Darby is designed to measure reading rate and retention at the high school or adult level. The reader is given a one-page practice test with 20 multiple choice practice questions and then five similar scored passages and sets of questions. Each test item consists of a section of the text. The reader is asked to mark the one phrase or sentence where the meaning of the original passage has been altered. Three scores are given for the total pool of 100 questions: rate (number of answers given), efficiency (number of correct answers), and accuracy (efficiency divided by rate, times 100). The test has one level, for which two alternate forms are sold.

PRICES

Reusable test booklets cost 50¢ each in sets of 30. Answer sheets are 6¢ each by the hundred, and scoring templates 50¢ each. Individual pupil reports in sets of 100 are 8¢ each. The manual, with technical data and directions for administration, costs $4.00. A specimen set is offered at $6.00. Date of information: 1977.

FIELD TEST DATA

After a developmental field test on 60 college students, validation and reliability studies were carried out with 61 and 41 college students, respectively.

ADMINISTRATION

The test is administered by an examiner to groups under timed conditions. Administration time is 25 minutes.

SCORING

A hand-scoring stencil is available.

COMMENTS

The manual includes a detailed discussion of the review of the Carver-Darby in Buros *Seventh Mental Measurements Yearbook*. Unlike the other tests reviewed here, the Carver-Darby is not built around instructional objectives. The design of the task is described in enough detail for the test to merit consideration as a criterion-referenced test, but our evaluative framework does not clearly fit the design of this unique test. The author intends to let the test go out of print when existing supplies are sold out.

CARVER-DARBY CHUNKED READING TEST Revrac Publications, 1970

MEASUREMENT PROPERTIES

A (B) C 1. Description. The construction and rationale of the test are well described. The authors state candidly that writing items for it is an art.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. Portions of text were selected at random to develop into test items, but it is not clear what population of information or skill the correct answers represent.

A (C) 4. Sensitivity. No data.

(A) C 5. Item Uniformity. Alternate form reliability is in the .7 to .8 range, all items presumably testing the same thing.

(A) C 6. Divergent Validity. Factor analysis shows a distinction between the rate and accuracy scores.

A (C) 7. Bias. No data.

(A) C 8. Consistency. Alternate form reliability in the range of .65 to .81 is reported for the three subscores.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A B C 15. Flexibility. Not relevant.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. Hand scoring with a template.

(A) C 19. Record Keeping. Individual pupil report sheets are available.

A (C) 20. Decision Rules. Decision rules are given for categorizing readers into six types, but a solid argument for these types is not made.

A (C) 21. Comparative Data. Data on the performance of 143 college students are given as a standard of comparison.

COOPER-McGUIRE DIAGNOSTIC WORD ANALYSIS TEST

Croft Educational Services, Inc., 1972

DESCRIPTION

The Cooper-McGuire battery consists of diagnostic tests for primary to intermediate grades that measure the following categories of skills: phonetic analysis, structural analysis, and readiness. There are 32 objectives with an average of 15 items each. Item formats include multiple choice, oral response, and fill-ins. The test of each objective is printed on separate spirit masters for local duplication and scoring. Alternate forms of this battery are available. An optional curriculum index is offered.

PRICES

The book of spirit masters for one form of the tests costs $26.00. Prices per test per pupil will vary with the number of objectives tested and number of copies made from each spirit master. The administrator's manual, which contains scoring keys, costs $8.00. The test manual, with objectives and rationale, is $2.00. Class record charts are $2.00 each in sets of 20, and individual pupil record cards are 12¢ each in packs of 50. Cassettes for administering the tests are $29.00 per set. Transparent overlays for scoring are $89.00 per set. The price for the curriculum index is $49.00. Date of information: 1978.

ADMINISTRATION

Except for six individually administered objectives, the tests are designed for group administration by a teacher or for self-testing by cassette tape.

SCORING

Hand scoring is done with filled in pupil pages or with optional overlays.

COOPER-McGUIRE DIAGNOSTIC WORD ANALYSIS TEST

Croft Educational Services, Inc., 1972

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. A number of the tests lack sample items.

A (C) 10. Item Review.

A (C) 11. Visibility. Several tests are too crowded for lower primary children.

(A) C 12. Responding.

A (C) 13. Informativeness. Specimen sets with full sets of objectives are not offered.

(A) C 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility. The test for each objective is printed on separate spirit master.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. Overlays are optionally available at extra expense. Reduced pupils' answer pages are in the administrator's manual. Machine scoring would not be appropriate.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Given, but without support.

A (C) 21. Comparative Data.

CRITERION-REFERENCED TESTS OF BASIC READING AND COMPUTATIONAL SKILLS

Multi-Media Associates, 1975

DESCRIPTION

The EPIC battery consists of eight levels of tests of core skills in reading and math for grades K-6. Each level measures 25 objectives with 4 multiple choice items per objective. At each level, 15 of the objectives are in reading, 10 in math. The reading skills tested range from identifying letters and sequencing story pictures at the lowest level to study skills and interpretive comprehension at the highest. Math skills range from counting objects up through ratios and proportions.

PRICES

The reusable notebook with 25 answer sheets for individually testing pupils in grades K-2 costs $8.75 per grade. A test package for one level of EPIC at grades 3-6 costs $8.75 and includes reusable tests for 25 students, 25 machine and hand scorable answer sheets, an examiner's manual, an answer key, and an envelope for ordering machine scoring. In such packages the unit price per test is 35¢ per pupil. Specimen sets are $8.00 per level for K-2 and $2.00 for each upper level. Date of information: 1978.

FIELD TEST DATA

Each level was field tested on 17 to 47 pupils at that level in Tucson, Arizona. Some items were revised after the field test. For each item, difficulty levels are reported. Test-retest consistency is reported in terms of percent of response consistency at the item level and correlations at the level of the objective and the total test.

ADMINISTRATION

The lower three levels are designed for individual testing, the upper levels for group testing by the teacher. Estimated testing time per level is 30 minutes for reading and 40 minutes for math.

SCORING

Hand scoring may be done by key or template, or the answer sheets may be machine scored. The basic scoring service costs 80¢ per answer sheet and includes individual scores, group summary scores, school summaries, and district summaries. Learner needs assessment reports and classroom item analysis reports are also available.

COMMENTS

Publisher also offers a customized test development service.

CRITERION-REFERENCED TESTS OF BASIC READING AND COMPUTATIONAL SKILLS

Multi-Media Associates, 1975

MEASUREMENT PROPERTIES

A B (C) 1. Description. Although the objectives are numerous and fairly narrow, they are vaguely stated.

A (C) 2. Agreement.

A (C) 3. Representativeness.

A (C) 4. Sensitivity.

A (C) 5. Item Uniformity.

A (C) 6. Divergent Validity.

A (C) 7. Bias.

(A) C 8. Consistency. In small samples of pupils (17-47) the test-retest reliabilities at the level of the objective ranged from a median in the .40s at level 7 to a median in the .70s at levels 3 and 4.

APPROPRIATENESS AND USABILITY

A B (C) 9. Instructions. Instructions for the lower level tests are complex. Sample items not provided for each objective.

(A) C 10. Item Review. Items were revised on the basis of the field test.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. There is considerable carry over of objectives from level to level, but all objectives for a level are tested in one booklet.

A (C) 16. Alternate Forms.

A (C) 17. Administration. Vague.

(A) B C 18. Scoring. Templates and machine scoring options are available.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Local option is offered, but a rationale for passing scores is not provided.

A (C) 21. Comparative Data. Difficulty levels are provided for all items, but on very small, local samples.

CRITERION TEST OF BASIC SKILLS Academic Therapy Publications, 1976

DESCRIPTION

This battery contains individual tests of 18 reading objectives and 26 math objectives for diagnosing the basic skills of pupils in grades K-8. Reading objectives deal with letters (recognition, sounding, and writing), phonics, and sight words, there being an average of 13 items per objective. The math objectives deal with numbers and numerals, the four basic operations, money, time, supplying the missing symbol, fractions, decimals, and percents, there being an average of over six items per objective. Item formats for reading and math are oral and fill-in. The manual has 70 pages of suggested teaching activities and materials.

PRICES

The complete test package, which includes stimulus cards, 25 answer sheets for each of reading and math, the administrator's manual, and a pad of math problems sells for $17.00. Replacement answer sheets are 14¢ each in sets of 25. Date of information: 1977.

FIELD TEST DATA

Field testing is mentioned, but not described.

ADMINISTRATION

These tests are designed for administration to individuals by a teacher. Estimated testing time is 10-15 minutes for each of the six sections in reading and eleven sections in math.

SCORING

Scoring is done on the spot by circling correct responses and writing incorrect responses on the individual pupil record.

COMMENTS

A word list from the local text series is used for the sight word objective. Publisher feels that features 6 and 14 are not appropriate for evaluating this test.

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Sample items are lacking.

(A) C 10. Item Review. Revision on the basis of field test is reported.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing. There is an extensive activities guide, but no indexing of test materials.

(A) B C 15. Flexibility.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Three levels are identified, but without support.

A (C) 21. Comparative Data.

DESIGN FOR MATH SKILL DEVELOPMENT
by D. A. Kamp, et al.

NCS Educational Systems,
1975

DESCRIPTION

The Design for Math Skill Development is a seven-level system for instructional management in elementary math that is built around the following ten content strands: numeration and place value, addition and subtraction, multiplication and division, word problems involving the basic operations, fractions, geometry, measurement, money, time, and graphing. The number of objectives per level ranges from 14 at the first to 30 at the highest, objectives averaging eight multiple choice items each. Two alternate forms are available.

PRICES

Test booklets average 37¢ per pupil in packages of 35 for levels A-D (consumables) and $1.71 for levels E-G (reusable). Placement tests average 29¢ in packets of 35. Spirit masters for printing answer sheets are $3.00. Also available are the Teacher's Planning Guide for $4.25, Administrator Manual at $1.25 for each level, and a Teacher's Resource File for $21.00. Date of information: 1978.

SCORING

Scoring is by hand using keys in the examiner's manual.

ADMINISTRATION

The Design for Math Skill Development test is administered to groups or individuals. Testing time for an entire level will take from 60 minutes at the lowest level to 195 minutes at the upper end.

DESIGN FOR MATH SKILL DEVELOPMENT
by D. A. Kamp, et al.

NCS Educational Systems,
1975

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review.

A (C) 11. Visibility. Tests at the lowest levels are crowded.

A (C) 12. Responding. At the lowest levels, response spaces are small.

(A) C 13. Informativeness.

(A) C 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility. Separate test forms for each objective, with good carry over of objectives across levels.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. Scoring by answer key.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

DIAGNOSIS: AN INSTRUCTIONAL AID - MATHEMATICS

Science Research Associates, 1972-73

DESCRIPTION

SRA's Diagnosis-Mathematics is a two-level battery of diagnostic tests for grades 1-6 that measure objectives in the following skill areas: computation, sets and numeration, operations, problem solving, measurement, and geometry. At each level there is a survey test and a series of diagnostic probe tests. The survey for Level A has 95 items testing 24 skill categories, while the 24 corresponding probe tests average 15 items each and have an average of about two items per objective. At Level B, the survey has 157 items testing 32 skill categories, and the 32 corresponding probe tests average 15 items each and 2 items per objective. All items are multiple choice. Alternate forms of the survey tests are optionally available. Two diagnostic labs, one for grades 1-4 and the other for grades 3-6, are available separately. These include diagnostic tests and prescriptive guides to basal texts and supplementary materials.

PRICES

A complete kit for a level lists at $80.00-$87.50 (school price--$60.00-$65.50). The kit for each level contains 30 copies of the survey test and of all the probes, the teacher's handbook, a guide to texts and materials, scoring overlays (Level A) or keys (Level B), etc. All test materials are consumable except the Level B surveys. Alternate forms of the surveys are available in sets of 30 for 28¢ per pupil list. Specimen sets for each level are $9.60 list ($7.20 school price). Date of information: 1977.

ADMINISTRATION

Tests are designed for group administration by an examiner or, in come cases, individual administration by the pupil.

SCORING

Level A is hand scored with overlay keys and Level B with strip keys.

COMMENTS

The teacher's handbook for Level B cross indexes the test items on several widely used norm-referenced tests to the SRA-Diagnosis probes and to sections of the SRA guide to texts and materials. A revised edition is expected to be on the market by 1980.

DIAGNOSIS: AN INSTRUCTIONAL AID - MATHEMATICS

Science Research Associates, 1972-73

MEASUREMENT PROPERTIES

A (B) C 1. Description. Objectives are printed on the backs of the test forms.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Sample items are not provided in the probes.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

A (C) 13. Informativeness. The objectives are not completely listed in the specimen materials.

(A) C 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility.

(A) (C) 16. Alternate Forms. Alternate forms of the survey are sold. Probes come in only one form.

(A) C 17. Administration.

A (B) C 18. Scoring. Hand scoring is easy.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

DIAGNOSIS: AN INSTRUCTIONAL AID - READING

Science Research Associates, 1974

DESCRIPTION

SRA's Diagnosis-Reading is a two-level battery of diagnostic tests for grades 1-6 that measure objectives in the following skill areas: phonetic analysis, structural analysis, comprehension, vocabulary, study skills, and use of sources. Each level (A=grades 1-4, B=grades 3-6) has a survey test of over 60 items and a series of over 30 diagnostic probes, each with an average of 20 items. On the probes, the minimum and usual number of items per objective is 2, there being 306 objectives at Level A and 224 at Level B. Item formats include multiple choice, matching, fill-in, and ordering. Alternate forms of the survey tests are optionally available. The classroom kit includes a guide to texts and other instructional materials. Two diagnostic labs, one for grades 1-4 and the other for grades 3-6, are available separately. These include diagnostic tests and prescriptive guides to basal texts and supplementary materials.

PRICES

A complete kit for Level A with 25 copies of the survey and of each of the probe tests, a guide to texts and materials, the teacher's handbook, cassettes for the phonetic tests, etc., lists for $159.50 (school price $119.50). The Level B kit lists for $116.75 (school price--$87.50). The alternate form of the survey for each level lists at 58¢ per pupil in sets of 25. Specimen sets for each level list at $9.60 ($7.20 school price). Date of information: 1977.

ADMINISTRATION

The SRA Diagnosis-Reading tests are made for a variety of modes of administration. The surveys are administered to groups by a teacher; Level A phonetics probes are given by cassette tape; and many of the other probes are self-administered under teacher supervision.

SCORING

Survey tests are scored by hand with a key, while the probes are self-scoring.

DIAGNOSIS: AN INSTRUCTIONAL AID - READING

Science Research Associates, 1974

MEASUREMENT PROPERTIES

A (B) C 1. Description. Objectives are located on the inside of the self-scoring answer sheets.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions. The survey tests do not have sample items, but the probes do.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

A (C) 13. Informativeness. The objectives are not completely listed in the specimen materials.

(A) C 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. Each probe tests from 2 to 26 objectives.

(A) (C) 16. Alternate Forms. Alternate forms of the survey are available for pre- and posttesting. Probes come in one form.

(A) C 17. Administration.

A (B) C 18. Scoring. The surveys are scored with reduced pupil pages; probes are self-scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

DIAGNOSTIC MATHEMATICS INVENTORY CTB/McGraw-Hill, 1975
by John K. Gessel

DESCRIPTION

A revision of the earlier Prescriptive Mathematics Inventory, the DMI is a seven-level diagnostic testing system for grades 1.5 through 7.5 plus. The following 11 categories of skills are covered: pre-operational concepts, counting, matching, addition of single digits, addition of integers with more than 1 digit, subtraction of integers, missing addends and factors, sequences and inequalities, measurement, plane figures, and inverse and place value. The DMI has from 37 to 179 multiple choice items per level, with each item testing a separate objective. The number of choices per test item ranges from 5 to 10. For math skills at this broad level of description--"measurement," "subtraction of whole numbers with regrouping," "segments, lines, rays"--the number of skills per level varies from 11 to 39 and the number of items per skill from 2 to 8.

To support classroom instruction, the following optional materials are available: interim tests for monitoring pupils' progress during the year, learning activities guides, guides indexing the DMI to math text series, and guides to non-text teaching materials.

PRICES

Test books come in packages of 35 with an Examiner's Manual. At the lower three levels, hand scorable test books are 38¢-49¢ per pupil, and machine scorable ones are 67¢-80¢. At the upper four levels, the reusable test books are 54¢-61¢ per pupil. Machine scored answer sheets are 13¢-16¢ per pupil in sets of 50; hand scorable ones are 20¢-40¢ in sets of 25. Consumable practice exercises for leveling pupils before giving the diagnostic tests are offered for 11¢ per pupil in sets of 35. The Teacher's Guide, serving all levels, is $3.25. Examination kits are $5.50 per level, $16.00 for an all-level kit. Date of information: 1979.

FIELD TEST DATA

A technical report was in preparation on the DMI while this volume was in progress. It has point-biserials for individual items, KR-20s, test-retest reliabilities, and item-difficulty data. Only small pieces of these data were available for our review.

ADMINISTRATION

The DMI is a group test designed to be given by an examiner. Estimated testing time varies from two sessions of 40 minutes each at the lowest level to six sessions of 45 minutes each at the upper levels.

SCORING

Tests are machine scored at a cost of 70¢ to 97¢ per pupil. The basic service includes all of the following: responses to individual items, individuals' total scores, summaries by item of group responses, summaries by total scores of the group. For an additional cost of 15¢ and 25¢ per pupil respectively, group and individual diagnostic reports are available. Estimated norms are optionally available. Publisher says that the approximate time for returning scores to the user is 15 days from receipt.

DIAGNOSTIC MATHEMATICS INVENTORY
by John K. Gessel

CTB/McGraw-Hill, 1975

MEASUREMENT PROPERTIES

A B (C) 1. Description. Objectives for single items defeat the purpose of objectives --to describe skills rather than test items. At higher levels of item grouping, the "objectives" are vague.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data on broader "category" objectives. Not applicable to one-item objectives.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No evidence.

A (C) 8. Consistency. No data. See notes under Field Test Data on the facing page.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

(A) C 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. Core objectives are tested at several levels, but testing of individual category objectives is not practical with machine scoring. Optional interim tests provide more flexibility.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

(A) B C 18. Scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Rules are implied but not well supported.

(A) C 21. Comparative Data. See notes under Field Test Data and Scoring on the facing page.

DOREN DIAGNOSTIC READING TEST OF WORD RECOGNITION SKILLS by Margaret Doren

American Guidance Service, 1973
(Original copyright: 1956)

DESCRIPTION

The Doren is a group diagnostic test for children in the primary grades which covers the following skills: letter and word recognition, beginning and ending sounds, consonants and vowels, word roots, blending, rhyming, spelling, sight words, and guessing words in context. Items are both multiple choice and written. There are 33 objectives averaging 12 items per objective, for a total of 395 items. Items are both multiple choice and written.

PRICES

Consumable test booklets are 27¢ each in sets of 25 booklets. An overlay key is offered for $5.90, and the manual (dated 1973) is $2.35. The specimen set, at $3.00, includes a test booklet, manual, and class record sheet. Date of information: 1977.

FIELD TEST DATA

Total test scores are reported for a sample of approximately 40 pupils at each of levels 1-4. The recency of these data is not reported.

ADMINISTRATION

The test is administered by an examiner in a group setting. The catalog estimates total testing time to be one to three hours depending on class size and reading level.

SCORING

Scoring is done either with an optional template or a key in the manual.

COMMENTS

The test was first published in 1956, before the days of objectives-based testing, but it was not apparently developed as a norm-referenced test.

DOREN DIAGNOSTIC READING TEST OF WORD RECOGNITION SKILLS by Margaret Doren

American Guidance Service, 1973
(Original copyright: 1956)

MEASUREMENT PROPERTIES

A (B) C	1. Description.	
A (C)	2. Agreement.	
A (C)	3. Representativeness.	
A (C)	4. Sensitivity. No data.	
A (C)	5. Item Uniformity.	
A (C)	6. Divergent Validity.	
A (C)	7. Bias. No data.	
A (C)	8. Consistency. No data.	

APPROPRIATENESS AND USABILITY

(A) B C	9. Instructions.	
A (C)	10. Item Review.	
(A) C	11. Visibility.	
(A) C	12. Responding.	
(A) C	13. Informativeness.	
A (C)	14. Curriculum Cross-Referencing.	The manual has a seven page section on remedial activities.
A B (C)	15. Flexibility.	
A (C)	16. Alternate Forms.	
(A) C	17. Administration.	
A (B) C	18. Scoring.	
(A) C	19. Record Keeping.	
A (C)	20. Decision Rules.	Easy-to-use decision rules are given, but they need justification.
A (C)	21. Comparative Data.	Ranges of scores are given by grade level for grades 1-4, but the field test population is limited to "four midwest suburban school districts."

EARLY CHILDHOOD ASSESSMENT
by Robert Wendt and Robert Schramm

Cooperative Educational Service, undated

DESCRIPTION

The ECA is a battery of performance tasks to be used for locating children of 3 to 6 years along a developmental curriculum sequence. It has six levels, ranging from sensory-motor activities which are maturationally determined through integration, to symbolic activities of reading and math. There are 73 separately scored objectives with the number of tasks for each ranging from one to twelve. The median number of items per objective for the 23 reading and 17 math objectives is 4 and 3 respectively. The manual states that the assessment is not designed to be diagnostic or categorical, but rather to serve as an aid for locating the child's level. A prescriptive guide to activities for learning centers in early childhood education is available separately.

PRICES

Consumable scoring booklets are 50¢ each, available in any numbers. The administrator's manual is $2.25 and the prescriptive guide is $4.75. Date of information: 1977.

FIELD TEST DATA

Field testing is mentioned, but not described.

ADMINISTRATION

The ECA is administered or supervised by a person trained in individual testing. Although it is an individual test, procedures are described for testing larger numbers of children at the same time at a series of separate testing stations. The following equipment is optional: an audiometer, Telebinocular, Titmus Stereo apparatus, Good-lite Screening Instrument. Estimated testing time is 45-50 minutes per pupil.

SCORING

Many of the objectives are scored by observer's judgment. Guidelines for scoring are often subjective.

COMMENTS

The ECA is an ESEA Title III Project. A new edition is scheduled to be published before the release of this handbook.

EARLY CHILDHOOD ASSESSMENT
by Robert Wendt and Robert Schramm

Cooperative Educational Service, undated

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Sample tasks are not consistently given.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding. Generally an action by the child.

A (C) 13. Informativeness. Program descriptors and an overview booklet are available on request at no cost.

A (C) 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. Parts of the test may be given to any one child.

A (C) 16. Alternate Forms.

A (C) 17. Administration. Instructions vary in clarity. Examiner has to switch between manual and detailed response booklet.

A B (C) 18. Scoring. Scoring will vary with the observer's subjective standards of correctness.

(A) C 19. Record Keeping. Pupil and class record sheets are provided.

A (C) 20. Decision Rules. Given but not with any support.

A (C) 21. Comparative Data.

EVERYDAY SKILLS TESTS: READING, TEST A; MATHEMATICS, TEST A

CTB/McGraw-Hill, 1975

DESCRIPTION

The Everyday Skills Tests consist of a battery of two objectives-based tests (Tests A) in the reading and math skills that are useful for adults in their daily lives and two norm-referenced tests (Tests B) in computation and reference/graphic materials. There are 3 multiple choice items for each of 15 reading objectives and 9 math objectives in the A tests. Reading objectives deal with materials like labels, ingredients, want ads, tax forms, and the like. Math objectives deal with matters like cost comparisons, rates of interest, and time calculations.

PRICES

Reusable test booklets are 32¢ each for reading and 26¢ each for math, both in sets of 35. The examiner's manual is included. Booklets contain both the objectives-based A part and norm-referenced B part of each domain. Answer sheets are 9¢ each in sets of 50 and scoring stencils are $2.75 apiece. A specimen set is offered at $5.00. Date of information: 1978.

FIELD TEST DATA

The A Tests were field tested in a sample of schools in Florida. Difficulties are reported for each item for 6th, 8th, and 10th grade pupils in the sample. The median percent of correct responses to an item at the 10th grade level is 88 in reading and 67 in math.

ADMINISTRATION

These tests are designed for group administration by a teacher. Estimated testing time is at least 30 minutes for reading and 24 minutes for math, although the tests are untimed. The norm-referenced parts of the battery, which are timed, take another 30-40 minutes each.

SCORING

Scoring is done by hand key, optional stencil, or machine. The scoring service provides a class record and individual reports for 50¢ and 75¢ per pupil, respectively.

COMMENTS

Items for part B of each test come from Form R of the Comprehensive Test of Basic Skills.

EVERYDAY SKILLS TEST: READING, TEST A; MATHEMATICS, TEST A

CTB/McGraw-Hill, 1975

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. Three items per objective were selected from a set of five partly on the basis of inter-item correlations, but the correlations are not given.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A B (C) 15. Flexibility.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

(A) B C 18. Scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data. Item difficulties are reported but on a sample of only 200-435 pupils from Florida.

FOUNTAIN VALLEY TEACHER SUPPORT SYSTEM IN MATHEMATICS

Richard L. Zweig Associates, 1972

DESCRIPTION

The Fountain Valley math tests are part of a nine-level diagnostic/prescriptive system which covers objectives for grades K-8 in the following areas: numbers and operations, geometry, measurement, applications, statistics and probability, sets, functions and graphs, logical thinking, and problem solving. The number of objectives per level ranges from 36 at the lowest to 135 at grade 6. Each test form contains the tests of several objectives, so there are 11 to 31 separate forms per level. The number of multiple choice items per objective ranges from two to twelve, with the average being at least three at all levels. Directions for all tests are given by cassette tape. An optional "teaching alternatives supplement" at each test level cross references the Fountain Valley objectives to the text and non-text instructional materials of 40 publishers.

PRICES

The total system for all grades (about 500 students) sells for about $2,750. An inservice (training) module is offered for $75.00. Modules containing a manual, a teaching alternative supplement, and tape cassettes sell for from $83.50 to $203 depending on the level. Hand-scored test forms are 3¢ per pupil in sets of 50, while the self-scoring forms are about 11¢ each. Answer keys sell for from $11 to $31 per level depending on the number of tests for the level. Rather than have the system described fully in a catalog or specimen set, the publisher explains the system mostly through its sale representatives. Date of information: 1977.

ADMINISTRATION

The Fountain Valley math tests are administered to groups of pupils for the most part by cassette tape. Administration of the tests of numbers and operations by teachers is supported by a separate manual. The estimated testing time per test form ranges from six to twenty minutes.

SCORING

Overlays are provided for scoring the answer sheets.

COMMENTS

The keying of items to objectives is contained only in the scoring materials, not with the objectives in the manuals.

FOUNTAIN VALLEY TEACHER SUPPORT SYSTEM IN MATHEMATICS

Richard L. Zweig Associates, 1972

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Sample items are not provided. The instructions at levels K and 1 may be too complex.

A (C) 10. Item Review.

(A) C 11. Visibility.

A (C) 12. Responding. Answer sheets for the lowest two levels are crowded.

A (C) 13. Informativeness.

(A) C 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. The test forms cover an average of three or four objectives each.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Rules are provided without support.

A (C) 21. Comparative Data.

FOUNTAIN VALLEY TEACHER SUPPORT SYSTEM IN READING — Richard L. Zweig Associates, 1975

DESCRIPTION

The Fountain Valley reading tests are part of a six-level system for the management of reading instruction in grades K through 6 covering the following skill areas: phonic analysis, structural analysis, vocabulary development, comprehension, and study skills. The number of objectives varies from 125 at level K-1 to 33 at level 4. Each test form contains the items for several (i.e., 3 to 6, on the average) objectives. There are from two to twelve multiple choice items per objective with the average being about three items at all levels. A "teaching alternatives supplement" cross references the tests' objectives to the text and non-text instructional materials of over 70 publishers.

PRICES

The total system for all grades (about 500 students) sells for about $2,125. An inservice (training) module is offered for $75.00. Modules containing the manual, the teaching alternatives supplement, and tape cassettes vary from $100 to $51 per level. Hand-scored test forms are 3.5¢ per pupil in sets of 50, while the self-scoring forms are about 12¢ each. Answer keys sell for from $9 to $19 per level depending on the number of tests for the level. Rather than have the system described fully in a catalog or specimen set, the publisher explains the system mostly through its sale representatives. Date of information: 1977.

FIELD TEST DATA

Over 10,000 students in grades 106 took part in the field test in the Fountain Valley, California, School District. Results of the field test are not reported.

ADMINISTRATION

These are group tests which can be administered by cassette tape or orally by a teacher.

SCORING

Two scoring options are offered: hand scoring by template or self-scoring with special answer sheets. Estimated testing time per test form is from 5.5 to 20 minutes.

COMMENTS

The keying of items to objectives is contained only in the scoring materials, not with the objectives in the manuals.

FOUNTAIN VALLEY TEACHER SUPPORT SYSTEM IN READING

Richard L. Zweig Associates, 1975

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement.

A (C) 3. Representativeness. Items were chosen according to how well they discriminated high and low scorers.

A (C) 4. Sensitivity. The data that are given in a mimeographed technical report indicate changes in scores on standardized tests following introduction of the system.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding. The answer sheets for the lower two levels are somewhat crowded.

A (C) 13. Informativeness. Specimen sets are not offered. It is hard to figure out what the basic package is.

(A) C 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. Test forms are one page each and test 3 to 6 objectives.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. Two methods of easy local scoring are offered: template and self-scoring answer sheet.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Provided, but without support.

A (C) 21. Comparative Data.

DESCRIPTION

The Group Phonics Analysis Test is a 75-item diagnostic test of basic phonics skills for pupils in grades 1-3. The 11 stated objectives range from recognizing printed letters and numbers to dividing words into syllables. There are from 3 to 19 multiple choice items per objective, the mode being 3.

PRICES

The one-page consumable test forms are 17¢ each in packs of 40, which include the examiner's manual. The specimen set, at $2.95, contains an examiner's manual and sample test as well as samples of four other tests by Dreier. Date of information: 1978.

FIELD TEST DATA

Norms and reliabilities are based on a field test of 104 pupils in grades 1-3.

ADMINISTRATION

These untimed tests are designed for group administration.

SCORING

The answer sheet contains a pressure-sensitive self-scoring second page.

MEASUREMENT PROPERTIES

A B (C) 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

(A) C 5. Item Uniformity. The KR-21 reliability for a sample of 104 pupils in grades 1-3 is .88.

(A) C 6. Divergent Validity. Scores on this test have a low correlation with scores on a test of reading comprehension (r=.32).

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Only two sample items are given.

A (C) 10. Item Review. No report.

A (C) 11. Visibility. See #12.

A (C) 12. Responding. The self-scoring test form is too crowded for easy answering by primary level children.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A B (C) 15. Flexibility.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. The test form has a pressure-sensitive self-scoring duplicate backing.

A (C) 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data. Inter-quartile bands are given around the average scores for 1st through 6th grades. The norming population is small and the method for inferring norms for grades 4-6 not described.

INDIVIDUAL PUPIL MONITORING SYSTEM - Mathematics — Houghton Mifflin, 1973-1974

DESCRIPTION

The IPMS-Mathematics is an eight-level system for continuously monitoring pupils' mastery of math objectives. The levels, corresponding roughly to grades 1-8, are each divided into three "assessment modules" aiming at one-third of a year of instruction. Tests for each objective are printed on separate pages. The number of objectives ranges from 48 at Level 1 to 64 at Level 8, the lower three levels having 5 multiple choice items per objective and the upper levels having 10. Two forms of the tests are available. In addition to the basic testing materials, resources for relating tests to instruction are optionally available, including one booklet at each level indexed to major math text series and guides to other learning materials and activities.

PRICES

Test booklets and individual pupil progress records come together in sets of 35 @ 48¢ per pupil, per module, per form for Level 1 and @ 57¢ for the other levels. Self-scoring answer sheets for Levels 3-8 sell for 13¢ each in sets of 100, and test booklets for those levels are reusable. The crayon for the self-scoring system is sold by the dozen at about 36¢ each. A Teacher's Kit containing a booklet of objectives, a set of classroom record forms, a teacher's guide, and a booklet indexing objectives to texts and teaching materials is available @$10.80 for Level 1 and $5.40 for the other levels. Date of information: 1978.

FIELD TEST DATA

The publisher reports field testing each level of each form on a national sample of about 350 pupils for the purpose of leveling and selecting test items from a larger initial pool of items.

ADMINISTRATION

Directions for group administration are provided, but at the upper levels pupils may be taking different tests at the same time. Tests are untimed, and no time estimates are provided. Pupils may be tested on as little as one objective at a sitting.

SCORING

Self-scoring by means of a latent-image system can be used, or scoring can be done by template.

COMMENTS

Several adults who tested the latent-image answer sheet and crayon found that heavy hand pressure was needed to make the hidden answer appear.

INDIVIDUAL PUPIL MONITORING SYSTEM - Mathematics

Houghton Mifflin, 1973-1974

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. A review is mentioned and the reviewers named, but the method is not described.

A (C) 3. Representativeness. An item analysis is mentioned but not described.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Instructions for higher levels are clear, but for lower level math concept items the vocabulary may be somewhat hard.

(A) C 10. Item Review. Item selection was based in part on field test data.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness. An informative specimen set is offered, but its contents should be described in the catalog.

(A) C 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

(A) C 16. Alternate Forms

(A) C 17. Administration.

A (B) C 18. Scoring. Two hand-scoring options are available.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

INDIVIDUAL PUPIL MONITORING SYSTEM - Reading — Houghton Mifflin, 1974

DESCRIPTION

The IPMS-Reading is a six-level system for managing instruction in reading. Designed for grades 1-6, it has from 43 to 63 objectives per level, each with five multiple choice test items. Tests for each objective are printed on separate pages. Two alternate forms of the tests are available. At every level there is one test booklet for each of these three groups of skills: word attack, vocabulary/comprehension, and discrimination/study skills. In addition to the basic testing materials, resources for relating tests to instruction are available, including indexes relating subtests to basal reading series (optional) and record keeping systems (included).

PRICES

Test booklets are 57¢ to 60¢ each, including an individual pupil record, in sets of 35. Booklets for the lower two levels are consumable. Self-scoring answer sheets for levels 3-6 are 13¢ each in sets of 100. Crayons for the self-scoring system are 36¢ each in sets of a dozen. Hand-scored answer sheets are about 5¢ in sets of 500. Teacher Kits @ $4.11 to $4.26 contain the following, which are also available separately: booklet of IPMS-Reading Objectives, Teacher's Guides, and Teacher's Management Record Booklet. For each of eight basal reading series, a separate cross-reference booklet is sold @ $1.98 to $3.30 per copy. The examination kit is $4.26. Date of information: 1978.

FIELD TEST DATA

A developmental field test is mentioned, but not described in any detail.

ADMINISTRATION

IPMS-Reading is administered to groups by an examiner. Time to administer these unspeeded tests will vary with the number of objectives tested at one sitting.

SCORING

Scoring can be done either by referring to answer keys at the back of the teacher's guide or by counting the correct items on the latent-image answer sheet.

COMMENTS

Several adults who tested the latent-image answer sheet and crayon found that heavy hand pressure was needed to make the hidden answer appear.

INDIVIDUAL PUPIL MONITORING SYSTEM - Reading — Houghton Mifflin, 1974

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. A review is mentioned, but not described.

A (C) 3. Representativeness. No information.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Sample items are included, but instructions for the lower levels appear too advanced in places.

A (C) 10. Item Review. No data.

A (C) 11. Visibility. At the lower levels items are crowded and response spaces are too small.

(A) C 12. Responding.

(A) C 13. Informativeness. An informative examination kit is offered, but its contents should be described in the catalog.

(A) C 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. Two methods of hand scoring are offered.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

INDIVIDUALIZED CRITERION-REFERENCED TESTING - Math	Educational Progress, 1973

DESCRIPTION

The ICRT-Math is an eight-level battery of tests for math in grades 1-8. At each level, there are 4 or 5 separate test forms of 16 multiple choice items (8 objectives) each. Objectives deal with sets, numeration systems, the four basic operations, geometry, functions and graphs, applications, and measurement. Alternate forms of the battery are available. Indexing of test objectives to two curriculum series by the publisher is given in the manual. Other prescriptive resources are optionally available.

PRICES

Test booklets are sold in complete sets for a level. In packages of 10 pupils, the price per pupil per test booklet runs 25¢ to 32¢. Machine scorable test forms are available for level 1; otherwise test forms are reusable in conjunction with answer sheets. Answer cards (plus the machine scoring service) are $1.25 per pupil for an order of at least 100 pupils. The teacher's/administrator's manual is $4.50. Date of information: 1976.

FIELD TEST DATA

The two ICRT components, math and reading, were field tested in six districts in Orange County, California. Data from the field tests are not reported.

ADMINISTRATION

The ICRTs-Math are made for group administration by a teacher or self-administration in the upper grades.

SCORING

Templates for hand scoring are available, and machine scoring is offered for 85¢ to $1.25 per pupil. The basic scoring service, which requires a minimum order for 100 pupils, includes prescriptive reports for individuals, an instructional grouping report for the class, a building summary, and a district summary. Estimated turnaround from receipt of materials is seven days.

INDIVIDUALIZED CRITERION-REFERENCED TESTING - Math

Educational Progress, 1973

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Most of the test forms lack sample items.

A (C) 10. Item Review.

(A) C 11. Visibility.

A (C) 12. Responding. Spaces for marking answers on machine scorable cards are small and crowded.

A (C) 13. Informativeness. No specimen set.

(A) C 14. Curriculum Cross-Referencing. In the manual, test objectives are indexed to two of the publishers' series of materials. With machine scoring, three other publishers' materials will be indexed in the reports.

A (B) C 15. Flexibility. There is much carry over of objectives from level to level, but each test form has items for eight objectives.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

(A) B C 18. Scoring. Hand scoring by template or machine scoring are available.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Provided, but with little justification for individual objectives. With only two items per objective, secure decisions about mastery of objectives are not possible.

A (C) 21. Comparative Data.

INDIVIDUALIZED CRITERION-REFERENCED TESTING - Reading	Educational Progress, 1973

DESCRIPTION

The ICRT-Reading tests make up an eight-level battery for pupils in grades K-8. The skills tested include word attack, literal comprehension, and interpretative comprehension. For levels 1-8, each test booklet has 16 multiple choice items, two questions for each of 8 objectives. The number of test booklets per level ranges from nine at level 1 to four at levels 4, 5, 6, and 7-8. At the K level, there are eight booklets of five items each. Indexing of test objectives to two curriculum series of the publisher is given in the manual. Other prescriptive resources are optionally available. Alternate forms of this battery are available.

PRICES

The package of 10 copies of all test booklets for one form of a level sells for $12.50, which gives a unit price of 32¢ or less per booklet. For levels 1-8, booklets are reusable. Answer sheets are $1.25 each for an order of at least 100, which includes the cost of machine scoring. Consumable booklets for levels 1 and 2 are also offered. In conjunction with the scoring materials, the unit price for these consumables is $2.85 per pupil for the entire level. A template and a 50-page answer sheet pad for local hand scoring cost $1.50 each per level for all levels. The tests for levels 1-8 are also packaged in a kit of 144 large cards for individual testing, one objective per side. This kit, called Benchmarks, is $38.50. Date of information: 1976.

FIELD TEST DATA

The two ICRT components, math and reading, were field tested in six districts in Orange County, California. Data from the field test are not reported.

ADMINISTRATION

Test booklets are made for group administration by a teacher. The tests as packaged in Benchmarks are for individual testing.

SCORING

Templates for hand scoring and machine scoring services are available. For a minimum order of 100 answer sheets, the cost is $125.00. It includes prescriptive reports for individuals, an instructional grouping report for the class, a building summary, and a district summary. Estimated turnaround time is seven days from receipt of materials.

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Sample items are lacking for most of the tests.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

A (C) 13. Informativeness. No specimen set is offered.

(A) C 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

(A) B C 18. Scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Provided, but with little justification for individual objectives. With only two items per objective, secure decisions about mastery of objectives are not possible.

A (C) 21. Comparative Data.

INSTANT WORD RECOGNITION TEST Dreier Educational Systems, 1971

DESCRIPTION

This test measures pupils' sight recognition of a 600-word basic 1st-4th grade vocabulary. It contains 48 multiple choice items that are arranged in increasing difficulty. An alternate form of this test is available by using a second orally presented word list with the same answer sheet.

PRICES

Self-scoring test forms are 17¢ each in sets of 30. The administrator's manual, which contains the 600-word basic vocabulary as well as directions for both forms, is included with an order of tests. A specimen set containing samples of this and several other tests by the publisher is available for $2.95. Date of information: 1978.

FIELD TEST DATA

On the basis of a field test of 153 first graders, a mean score of 11.1 correct and a correlation of +.77 with a standardized test is reported.

ADMINISTRATION

Group administration by a teacher is intended.

SCORING

The pupils' answer sheets are self-scoring.

COMMENTS

No objective is stated as such, but the criterion pool of words is given in the manual.

MEASUREMENT PROPERTIES

A B (C) 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review.

A (C) 11. Visibility. There are too many items per page for first graders.

(A) C 12. Responding. Except for #11 above.

(A) C 13. Informativeness. The promotional and sample materials tell what the test is like. But on looking at the test and manual, it is hard to know what a test score means.

A (C) 14. Curriculum Cross-Referencing.

A B (C) 15. Flexibility.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring.

A (C) 19. Record Keeping. Only spaces for raw scores on the answer sheet are provided.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data. The one average score given is based on a small sample.

KEYMATH DIAGNOSTIC ARITHMETIC TEST
by Austin J. Connolly, William Nachtman, & E. Milo Pritchett

American Guidance Service, 1971
(Metric Supplement, 1976)

DESCRIPTION

KeyMath is a diagnostic battery intended for individually testing pupils in grades K-6. At the level of specific objectives, there are 209 objectives, each with one test item. Objectives (items) are grouped into subtests as follows: numeration, fractions, geometry and symbols, the four basic operations, mental computation, numerical reasoning, word problems, missing elements, money measurement, and time. Subtests have 7 to 27 items that are arranged on a scale of progressive difficulty as determined by Rasch-Wright item analysis methods. Within subtests, items are grouped into "instructional clusters" of an average of 2 to 3 items. A 31-item metric supplement is also offered.

PRICES

A complete Test Kit is $26.50. The price of each component item, if ordered separately, is as follows: Reusable Easel-Kit at $21.50, examiner's manual at $2.85, and Diagnostic Records per package of 25 at $4.55. The Metric Supplement Manual and Test items sell for $4.25, and the response forms for it are $2.50 per package of 25. Date of information: 1978.

FIELD TEST DATA

Over 2000 pupils in a national sample were field tested, 1222 of them for norming KeyMath. Grade equivalents and W-scale values for total scores and for each individual item are given.

ADMINISTRATION

KeyMath is made for individual testing. Estimated testing time is 30 minutes per pupil.

SCORING

Stimulus pictures have correct answers printed on the flip side for immediate scoring and recording.

COMMENTS

Publisher says that the test may be used for remedial purposes above grade 6. Fall and spring percentile norms and normal curve equivalents for KeyMath were expected to be available by the time this volume is published.

KEYMATH DIAGNOSTIC ARITHMETIC TEST
by Austin J. Connolly, William Nachtman, & E. Milo Pritchett

American Guidance Service, 1971
(Metric Supplement, 1976)

MEASUREMENT PROPERTIES

A B (C) 1. Description. There are objectives for individual items, but not for clusters of items.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness.

A (C) 4. Sensitivity.

A (C) 5. Item Uniformity. Split-half reliabilities for the 14 subtests range from .23 to .90 within grade with the median (of all grades) ranging from .64 to .84.

A (C) 6. Divergent Validity.

A (C) 7. Bias.

A (C) 8. Consistency. The publisher advises against interpreting its test-retest data as reliabilities owing to the long period which separated the two testings.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Sample items are not given.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness. The system is available on approval.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. Machine scoring is not relevant here. Scoring is done on the spot with pre-printed answers on backs of question cards.

A (C) 19. Record Keeping. A graphic profile record is provided, but it is keyed to subtests and to individual items, not to instructional clusters.

A (C) 20. Decision Rules. Provided, but without support. Decisions are for subtests, not for "instructional clusters" of items.

A (C) 21. Comparative Data. The 1971 norms are given in grade equivalents. Only five school districts took part in the calibration study.

LANGUAGE AND THINKING PROGRAM: MASTERY LEARNING CRITERION TESTS

Follett Publishing Company, 1973

DESCRIPTION

The language and thinking tests measure children's proficiency in selecting pictures of familiar things in response to different categories of verbal instructions. The publisher says that the tests may be used for mastery testing or regrouping. Separate test booklets are provided for each of these groups of verbal concepts: classification, functions, directions/locations, colors/shapes/sizes, actions, and blends (i.e., combinations of two or more features). Designed for children from 3 to 7 years, the tests are almost entirely multiple choice. Each test booklet measures from 6 to 12 objectives, the number of items per objective ranging from two to eight.

PRICES

Consumable test booklets cost from 36¢ to 60¢ each, or $3.42 for the set of 7 (six concept areas plus a practice booklet). Reusable examiner's manuals for each test are from $1.14 to $1.83 each, or $9.96 for the set. Date of information: 1977.

FIELD TEST DATA

Data are not reported, but the commercially available edition of the test that was reviewed by CSE was the field research edition.

ADMINISTRATION

These are group tests which are given by an examiner.

SCORING

Scoring is by hand from keys in the examiner's manuals.

COMMENTS

These tests were developed as part of the Language and Thinking Program of CEMREL, Inc., but are sold separately.

LANGUAGE AND THINKING PROGRAM: MASTERY LEARNING OF CRITERION TESTS

Follett Publishing Company, 1973

MEASUREMENT PROPERTIES

A B (C) 1. Description. Several test objectives reflect two or more instructional objectives.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A B (C) 9. Instructions. The language of the instructions is advanced for pupils of this age.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing. Tests are keyed to the specific language and thinking instructional package with which they were developed.

A (B) C 15. Flexibility. The different concept areas may be tested separately, but there is only one level for each.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data. Not available for the Field Research Edition.

LANGUAGE ARTS: COMPOSITION, LIBRARY, AND LITERARY SKILLS (K-6)

Instructional Objectives Exchange, 1973

DESCRIPTION

This battery is a collection of multiple choice and fill-in tests measuring 16 objectives on composition, 10 on library skills, and 6 on literary skills. There are from five to ten items per objective. Tests for each objective are printed on spirit masters for local duplication and scoring. Two alternate forms of this collection are available.

PRICES

Each form of this test collection sells for $29.95, which includes the manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used. Date of information: 1979.

FIELD TEST DATA

Publisher reports that each test was tried out on at least five students in an elementary school in Los Angeles. Data are not reported.

ADMINISTRATION

These are group tests.

SCORING

Answer keys are provided in the manual for hand scoring.

COMMENTS

Leveling of the tests in this collection according to content, format, etc., is done locally by the test user. The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C 1. Description. Amplified objectives: rules for sampling each domain are not given though.

A (C) 2. Agreement. A review is reported but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. The one-page scoring guide contains keys for all 32 tests in small print.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

LANGUAGE ARTS: MECHANICS AND USAGE (K-6)

Instructional Objectives Exchange, 1973

DESCRIPTION

This battery is a collection of fill-in and multiple choice tests measuring 10 objectives in mechanics (capitalization and punctuation) and 23 objectives in usage (plurals, possessives, modifiers, verb agreement, irregular verbs, and commonly confused words). There is an average of more than eight items per objective. Tests for each objective are printed on spirit masters for local duplication and scoring. Two alternate forms of this collection are available.

PRICES

Each form of this test collection sells for $29.95, which includes the manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used
Date of information: 1979.

FIELD TEST DATA

Publisher reports that each test was tried out on at least five students in an elementary school in Los Angeles. Data are not reported.

ADMINISTRATION

These are group tests.

SCORING

Answer keys are provided in the manual for hand scoring.

COMMENTS

Leveling of the tests in this collection according to content, format, etc., is done locally by the test user. The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C	1. Description. Amplified objectives: rules for sampling each domain are not given though.
A (C)	2. Agreement. A review is reported, but not described.
A (C)	3. Representativeness. No data.
A (C)	4. Sensitivity. No data.
A (C)	5. Item Uniformity. No data.
A (C)	6. Divergent Validity. No data.
A (C)	7. Bias. No data.
A (C)	8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C	9. Instructions.
(A) C	10. Item Review.
(A) C	11. Visibility.
(A) C	12. Responding.
(A) C	13. Informativeness.
A (C)	14. Curriculum Cross-Referencing.
(A) B C	15. Flexibility.
(A) C	16. Alternate Forms.
(A) C	17. Administration
A B (C)	18. Scoring. Keys for all 33 objectives are printed on two pages in small type.
(A) C	19. Record Keeping.
A (C)	20. Decision Rules.
A (C)	21. Comparative Data.

LANGUAGE ARTS: WORD FORMS AND SYNTAX (K-6) — Instructional Objectives Exchange, 1973

DESCRIPTION

This battery is a collection of selected response tests measuring 15 objectives dealing with word form and 27 objectives dealing with syntax. There are five to ten items per objective. Tests for each objective are printed on spirit masters for local duplication and scoring. Two alternate forms of this collection are available.

PRICES

Each form of this test collection sells for $29.95, which includes the manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used. Date of information: 1979.

FIELD TEST DATA

Publisher reports that each test was tried out on at least five pupils in an elementary school in Los Angeles. Data are not reported.

ADMINISTRATION

These are group tests.

SCORING

Answer keys are provided in the manual for hand scoring.

COMMENTS

Leveling of these tests according to content, format, etc., is done locally by the user. The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C 1. Description. Amplified objectives: rules for sampling each domain are not given though.

A (C) 2. Agreement. A review is mentioned but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. Keys for all 42 objectives are printed on three pages in small type.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

MASTERY: AN EVALUATION TOOL (MATHEMATICS)

Science Research Associates, 1974-75

DESCRIPTION

Mastery (Math) is a nine-level battery of tests in math for grades K-8. There are 15 to 40 objectives per level with three multiple choice items per objective. The following skill areas are covered by the catalog (that is, ready-made) tests: for K-2--numbers and numerals, whole-number computation, measurement, sets, logical thinking, and geometry; for 3-8--whole numbers, fractional numbers, integers, rational and real numbers, geometry, measurement, sets, functions, graphing, statistics, probability, logic, and flow charts. Two alternate forms are available.

PRICES

Test booklets are 55¢ to 79¢ each per level in sets of 25, the lower three levels being consumable. Answer sheets are 13¢ each in sets of 100. An examiner's manual, which is included with an order of tests, is available separately for 70¢. Catalogues of Mastery (Math) objectives are available at $2.20 for the K-2 set and $3.55 for the 3-8 set. Specimen set for K-2 is $5.00 and for 3-9 it is $5.25. Date of information: 1977.

FIELD TEST DATA

A technical report is available from SRA giving item difficulties, item/test correlations, and KR-20s for each test level. Data come from "a cross-section of SRA test users." Numbers of test takers average about 3000 per level for form X and 475 for form Y.

ADMINISTRATION

Mastery (Math) is a battery of group tests designed to be given by a teacher. Estimated testing time is three minutes per objective.

SCORING

Keys are provided for hand scoring and a machine scoring service is offered. For a price per pupil of 98¢ to $1.40 the user receives profiles for individual pupils and for the total group.

COMMENTS

The publisher offers a customized CRT service as well as catalog (ready-made) tests.

MASTERY: AN EVALUATION TOOL (MATHEMATICS)

Science Research Associates, 1974-75

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. A review of items for their content validity is mentioned, but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. Point biserial correlations of items with the total test score have a median of about .4, and the KR-20s for test levels have a median of .95.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review. A review is mentioned, but not described in any detail.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness. Tests are available on 30-day approval.

(A) C 14. Curriculum Cross-Referencing. Available separately.

A (B) C 15. Flexibility. Catalog tests cover similar objectives at several levels, but all objectives are in one booklet per level.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. Both machine and hand scoring are available, but hand scoring does not appear easy.

A (C) 19. Record Keeping. If the scoring service is purchased, detailed records are provided.

(A) C 20. Decision Rules. For each three-item objective, the probabilities of attaining scores of 0 to 3 by guessing are provided.

A (C) 21. Comparative Data. Although the samples for the item-difficulty data in the technical report are large, the publisher does not claim that they are necessarily representative of the nation.

MASTERY: AN EVALUATION TOOL (SOBAR READING) — Science Research Associates, 1975

DESCRIPTION

SOBAR (System for Objective Based Assessment of Reading) is a ten-level battery for testing the following reading skills in grades K-9: letter recognition, phonic analysis, structural analysis, vocabulary, comprehension, and study skills. There are three multiple choice items per objective, the number of objectives ranging from 23 at level K to 35 at the upper levels. Two alternate forms are available.

PRICES

In sets of 25, test booklets range from 79¢ per pupil for the lower three levels (consumable) to 55¢ for the upper levels (reusable). The examiner's manual, which comes with an order of test booklets, may be bought separately for 70¢ depending on the level. Answer sheets are 13¢ each in packages of 100. Catalogs of SOBAR objectives cost approximately $2.95 each, there being a K-2 and a 3-9 catalog. Specimen set for K-2 sells for $5.00 and for 3-9 it is $5.25. Date of information: 1977.

FIELD TEST DATA

A technical report is available from SRA giving difficulty statistics for each item, point-biserials for each item, and KR-20s for each test level. Numbers of test takers averaged about 3200 per level for form L and 450 per level for form M.

ADMINISTRATION

SOBAR is a battery of group tests to be administered by the teacher. Estimated testing time is three minutes per objective.

SCORING

Keys are provided for hand scoring, and a machine scoring service is offered. For a per pupil price of $.98-$1.40, the buyer receives profiles for individual pupils and for the group.

COMMENTS

In addition to the catalog (ready-made) tests, the publisher offers a customized CRT service.

MASTERY: AN EVALUATION TOOL (SOBAR READING)

Science Research Associates, 1975

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. A review of items for their congruence with their objectives is mentioned but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. Point biserial correlations of items with total test scores are reported, KR-20s for test levels have a median of .94.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. A review of the items for racial and sexual bias is mentioned but not described.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review. A review is mentioned but not described in any detail.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

(A) C 14. Curriculum Cross-Referencing. Available separately.

A (B) C 15. Flexibility. Catalog (ready made) tests are in one booklet per level. Objectives are covered at several levels.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. Both machine and hand scoring are available, but hand scoring does not appear easy.

A (C) 19. Record Keeping. If the scoring service is purchased, detailed records are provided.

(A) C 20. Decision Rules. For each three-item objective, the probabilities of a pupil getting scores of 0-3 by guessing are provided.

A (C) 21. Comparative Data. Although the samples for the item-difficulty data in the technical report are large, the publisher does not claim that they are necessarily representative of the nation.

DESCRIPTION

The U-SAIL Math Tests make up a six-level battery for pupils in grades 1-6 on the following concepts: whole numbers, basic operations with integers, basic operations with fractions and decimals, sets measurement, geometry, graphs and functions, ratio and proportion, and percent. There are 10 to 17 objectives per level with five multiple choice items per objective. These tests are part of a math curriculum which includes instructional materials and other resources for teachers.

PRICES

Consumable tests for the lower three levels range from 24¢ to 37¢ per pupil in sets of 35, and reusable tests for the upper levels range from 22¢ to 24¢ per pupil in the same quantity. The teacher's manual is 75¢. A complete set of all 35 copies of all the levels is $56.00. Date of information: 1978.

FIELD TEST DATA

U-SAIL provided CSE with some unpublished data on item difficulties and inter-item correlations within each objective for the lower four levels. The number of pupils per item was 223 to 249. It is these data that are referred to below in the comments on standards 5 and 21. Test data were used for revision of the materials.

ADMINISTRATION

U-SAIL tests are designed for group administration.

SCORING

Templates for hand scoring are provided with the test booklets.

COMMENTS

This test battery was developed by a consortium of school districts.

MATH DIAGNOSTIC/PLACEMENT TESTS

U-SAIL (Utah System Approach to Individualized Learning Project), 1975

MEASUREMENT PROPERTIES

A B (C) 1. Description. For users of the U-SAIL math program, the ratings on test features #1-3 would be higher, since the items are systematically sampled from the domains that make up the curriculum. For the general test buyer, the scope and sequence chart gives only brief descriptions of the math objectives.

A (C) 2. Agreement.

A (C) 3. Representativeness.

A (C) 4. Sensitivity. The unpublished data of pupil gains are not clearly free from well-known problems in measurement.

(A) C 5. Item Uniformity. Part-whole correlations per objective are reported for the lower four levels. Most are in the .6 to .7 range.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. Unpublished information from the developer refers to studies to ensure lack of bias, but details are lacking.

A (C) 8. Consistency. The unpublished data provided by the developer were not complete enough to evaluate.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

A (C) 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing. Although the developer does not provide a curriculum index for these tests, it states that many publishers of math programs do index their text series to the U-SAIL objectives.

A (B) C 15. Flexibility. Each objective is covered at only one level, but the use of more than one level of test with individual pupils is suggested.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. By hand template.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Three levels of attainment are described, but not supported.

A (C) 21. Comparative Data. The publisher has some comparative data, but does not routinely provide them to test buyers. The pupils were from a geographically limited area.

MATHEMATICS: ELEMENTS, SYMBOLISM, AND MEASUREMENT (7-9)

Instructional Objectives Exchange, 1974

DESCRIPTION

This battery is a collection of multiple choice and fill-in tests dealing with 43 objectives in the following skill areas: integers, rational numbers, real numbers, numeration, measurement, and sentences and logic. The items for each objective are printed on separate spirit masters for local duplication and scoring. Number of items per objective ranges from 5 to 10. Two alternate forms of this collection are sold.

PRICES

Each form of this test collection sells for $29.95, which includes the manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used. Date of information: 1979.

FIELD TEST DATA

Preliminary field testing of these materials was done in two schools in Los Angeles.

ADMINISTRATION

These are group tests which may also be self-administered by pupils.

SCORING

Answer keys are provided in the manual for hand scoring.

COMMENTS

The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C 1. Description. Amplified objectives are given for all tests, but rules for sampling the domains are not.

A (C) 2. Agreement. Reviews of agreement are reported, but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility. The test for each objective is printed on a separate spirit master.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. The print is small and crowded on the answer keys.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

MATHEMATICS: GEOMETRY (K-6)

Instructional Objectives Exchange, 1973

DESCRIPTION

This battery is a collection of multiple choice and fill-in tests dealing with 36 geometry objectives. There are five items per objective, with the tests for each objective being printed on separate spirit masters for local duplication and scoring. Two alternate forms of this collection are available.

PRICES

Each form of this test collection sells for $29.95, which includes the manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used. Date of information: 1979.

FIELD TEST DATA

Preliminary field testing of these materials was done in two schools in Los Angeles.

ADMINISTRATION

These are group tests.

SCORING

Answer keys are provided in the manual for hand scoring.

COMMENTS

The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C 1. Description. Amplified objectives are given for all tests, but rules for sampling the domain are not.

A (C) 2. Agreement. Reviews of agreement are reported, but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A B (C) 15. Flexibility. The test for each objective is printed on a separate spirit master.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. The print is small and crowded on the answer keys.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

MATHEMATICS: GEOMETRY, OPERATIONS, AND RELATIONS (7-9)

Instructional Objectives Exchange, 1973

DESCRIPTION

This battery is a collection of multiple choice and fill-in tests covering 48 objectives in the following skill areas: geometry, operations and properties, statistics, ratios and proportions, and graphs. There are at least five items per objective, the tests for each objective being printed on separate spirit masters for local duplication and scoring. Two alternate forms of this collection are sold.

PRICES

Each form of this test collection sells for $29.95, which includes the manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used. Date of information: 1979.

FIELD TEST DATA

Preliminary field testing of these materials was done in two schools in Los Angeles.

ADMINISTRATION

These tests may be administered to groups by an examiner, and may be self-administered by pupils.

SCORING

Answer keys are provided in the manual for hand scoring.

COMMENTS

The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C 1. Description. Amplified objectives are given for all tests, but rules for sampling the domains are not.

A (C) 2. Agreement. Reviews of agreement are reported, but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility. The test for each objective is printed on a separate spirit master.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. The print is small and crowded on the answer keys.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

DESCRIPTION

This battery is a collection of multiple choice and fill-in tests covering 38 elementary level objectives in measurement. There are five items per objective, the test for each objective being printed on separate spirit masters for local duplication and scoring. Two alternate forms of this collection are available.

PRICES

Each form of this test collection sells for $29.95, which includes the manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used. Date of information: 1979.

FIELD TEST DATA

Preliminary field testing of these materials was done in two schools in Los Angeles.

ADMINISTRATION

These are group tests.

SCORING

Answer keys are provided in the manual for hand scoring.

COMMENTS

The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C 1. Description. Amplified objectives are given for all tests, but rules for sampling the domains are not.

A (C) 2. Agreement. Reviews of agreement are reported, but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility. The test for each objective is printed on a separate spirit master.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. The print is small and crowded on the answer keys.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

MATHEMATICS: NUMERATION AND RELATIONS (K-6)

Instructional Objectives Exchange, 1973

DESCRIPTION

This battery is a collection of multiple choice and fill-in tests covering 38 objectives in the following skill areas: numeration, ratios and proportions, graphs, statistics and probability, and logic. There are five to ten items per objective. The items for each objective are printed on separate spirit masters for local duplication and scoring. Two alternate forms of the collection are sold.

PRICES

Each form of this test collection sells for $29.95, which includes the manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used. Date of information: 1979.

FIELD TEST DATA

Preliminary field testing of the materials was done in two schools in Los Angeles.

ADMINISTRATION

These are group tests.

SCORING

Answer keys are provided in the manual for hand scoring.

COMMENTS

The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C 1. Description. Amplified objectives are given for all tests, but rules for sampling the domains are not.

A (C) 2. Agreement. Reviews of agreement are reported, but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATFNESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility. The test for each objective is printed on a separate spirit master.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. The print is small and crowded on the answer keys.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

MATHEMATICS: OPERATIONS AND PROPERTIES (K-6)

Instructional Objectives Exchange, 1974

DESCRIPTION

This battery is a collection of multiple choice and fill-in tests dealing with 40 objectives on the four basic operations--addition, subtraction, multiplication, and division--using integers, fractions, and decimals. There are five items per objective. The tests for each objective are printed on separate spirit masters for local duplication and scoring. Two alternate forms of this collection are available.

PRICES

Each form of this test collection sells for $29.95, which includes the manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used. Date of information: 1979.

FIELD TEST DATA

Preliminary field testing of the materials was done in two Los Angeles schools. After publication, performance data on 200 to 600 pupils per objective were gathered.

ADMINISTRATION

These are group tests.

SCORING

Answer keys are provided in the manual for hand scoring. Comparative data are given in the form of cumulative percentages of pupils at each of two to four grades attaining each possible score for each objective. Pupils were tested in the fall, so the publisher reports data for each group as year-end results for the previous grade.

COMMENTS

The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C 1. Description. Amplified objectives are given for all tests, but rules for sampling the domains are not.

A (C) 2. Agreement. Reviews of agreement are reported but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility. The test for each objective is printed on a separate spirit master.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. The print on the answer keys is small and crowded.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data. Data are provided, but the samples are not large and are all from urban settings in Southern California.

MATHEMATICS: SETS AND NUMBERS (K-6) Instructional Objectives Exchange, 1973

DESCRIPTION

This battery is a collection of multiple choice and fill-in tests dealing with 35 objectives in the following skill areas: sets, whole numbers, and rational numbers. There are five items per objective. Tests for each objective are printed on separate spirit masters for local duplication and scoring. Two alternate forms of this collection are sold.

PRICES

Each form of this test collection sells for $29.95 which includes the manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used. Date of information: 1979.

FIELD TEST DATA

Preliminary field testing of these materials was done in two schools in Los Angeles.

ADMINISTRATION

These are group tests.

SCORING

Answer keys are provided in the manual for hand scoring.

COMMENTS

The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C 1. Description. Amplified objectives are given for all tests, but rules for sampling the domains are not.

A (C) 2. Agreement. Reviews of agreement are reported but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility. The test for each objective is printed on separate spirit masters.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. The print is small and crowded on the answer keys.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

MCGUIRE-BUMPUS DIAGNOSTIC COMPREHENSION TEST

Croft Educational Services, 1971-72

DESCRIPTION

The McGuire-Bumpus tests are a two-level battery for primary and intermediate pupils which measure the following types of reading comprehension skills: literal, interpretive, analytic, and critical. The number of objectives for each of these skill types is respectively 4, 3, 3, and 2 at each level, each objective having 12 multiple choice items. Tests are printed on spirit masters for local duplication and scoring. Alternate forms are available. An optional curriculum index is offered.

PRICES

The book of spirit masters for one form of the tests costs $26.00. Prices per test per pupil will vary with the number of objectives tested and number of copies made from each spirit master. The administrator's manual, which contains scoring keys, costs $8.00. Scoring overlays may be ordered at $89.00 for one test form. Class record charts are $2.00 each in sets of 20, and individual pupil records are 12¢ each in sets of 50. Cassettes for administering the tests are $29.00 per set. The curriculum index sells for $49.00. Date of information: 1978.

ADMINISTRATION

These tests are made for group administration by a teacher or for self-administration by cassette recorder.

SCORING

Hand scoring is done with answer keys in the manual or with optional overlays.

COMMENTS

The test battery by itself lacks explanatory and interpretive information.

MCGUIRE-BUMPUS DIAGNOSTIC COMPREHENSION TEST

Croft Educational Services, 1971-72

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

(A) C 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. Each objective is tested at two levels, but individual objectives are not separately testable.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. Overlays are available. The keys in the manual are not so easy to use.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Rules are given without support.

A (C) 21. Comparative Data.

NEW MEXICO CAREER EDUCATION TEST SERIES by C. C. Healy & S. P. Klein

Monitor, 1973

DESCRIPTION

The New Mexico Career Education Test Series is a battery of tests dealing with career related attitudes, knowledge, and activities for pupils in grades 9-12. The four cognitive tests deal with these subjects: career planning, knowledge of occupations, job application procedures, and career development. Each of these tests has 20 to 25 multiple choice items divided among two or three sub-objectives. Two forms of the career planning test are offered.

PRICES

Reusable booklets for each of the tests are 24¢ per pupil in sets of 35. Answer sheets are 6¢ each in like sets. The examiner's manual for the series is $2.50 and separate answer keys are $1.00 per test. A specimen set is $3.75 for each test and $17.50 for the series. Date of information: 1978.

FIELD TEST DATA

Each of the tests was given to a sample of at least 500 ninth graders and 1200 twelfth graders in New Mexico. Item difficulties, point biserials, and norms are given for all tests.

ADMINISTRATION

These tests are designed to be given to groups. They are timed, taking 20 minutes each.

SCORING

Tests are scored by hand with templates.

COMMENTS

Eight of the items on the career development test measure an affective objective.

NEW MEXICO CAREER EDUCATION TEST SERIES by C. C. Healy & S. P. Klein

Monitor, 1973

MEASUREMENT PROPERTIES

A B (C) 1. Description.

A (C) 2. Agreement.

A (C) 3. Representativeness.

A (C) 4. Sensitivity. Small but statistically reliable differences in the scores of 9th and 12th graders are reported. Whether these differences are due to instruction cannot be determined from the data.

A (C) 5. Item Uniformity. Internal consistency measures range from .51 to .87 for the separate tests but the data are for total tests, not for the separate objectives. An average of five items per test have correlations with the total test score of less than .3.

A (C) 6. Divergent Validity.

A (C) 7. Bias.

(A) C 8. Consistency.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review.

A (C) 11. Visibility. Print size in the test items is small.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. The series has four separately sold components, each with 2-3 sub-objectives.

A (C) 16. Alternate Forms. Only one of the tests has two forms.

(A) C 17. Administration.

A (B) C 18. Scoring. By template.

A (C) 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data. Norming samples range from 500 to 2500 pupils, all from New Mexico.

NEW MEXICO CONCEPTS OF ECOLOGY TEST Monitor, 1973

DESCRIPTION

The Concepts of Ecology Tests are a two-level battery of survey tests in ecology for grades 6-12. Each level has 20 items and deals with 5 to 7 "knowledge areas." There are 2 to 6 multiple choice items per knowledge area.

PRICES

Reusable test booklets are 24¢ per pupil in sets of 35 and answer sheets are 6¢ each in like sets. The examiner's manual is $1.50 and answer keys are $1.00 per level. A specimen set is available at $3.00 per level. Date of information: 1978.

FIELD TEST DATA

The lower level was field tested on 1,040 sixth grade students, the upper level on 2,389 12th graders, both groups in New Mexico. Difficulties and other statistics are reported for each item, as are internal consistencies and norms for the whole test.

ADMINISTRATION

These tests are designed for group administration. They are timed, taking 20 minutes each.

SCORING

Scoring is done by hand with a template.

MEASUREMENT PROPERTIES

A B (C) 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness.

A (C) 4. Sensitivity. An average superiority of about two items correct for 12th graders over 9th graders is reported, but that gain is not clearly attributable to instruction.

A (C) 5. Item Uniformity. Internal consistencies of .67 and .74 are reported for the total test, but consistencies by "knowledge area" are not given. Four to five items per test have correlations with the total test score of less than .3.

A (C) 6. Divergent Validity.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A B C 15. Flexibility. Not clearly relevant. Four of the knowledge areas are tested on both levels.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. By hand with templates.

A (C) 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data. The norm samples are from one state, New Mexico.

NEW MEXICO CONSUMER MATHEMATICS TEST & CONSUMER RIGHTS AND RESPONSIBILITIES TEST — Monitor, 1973

DESCRIPTION

There are two New Mexico Consumer Tests, the Consumer Mathematics Test and the Consumer Rights and Responsibilities Test. Designed for pupils in grades 9-12, both contain 20 items. Clusters of generally three items deal with more specific topics such as insurance or unit prices.

PRICES

Reusable booklets for each test are 24¢ per pupil in sets of 35, and answer sheets are 6¢ in like sets. An examiner's manual for each test is $1.50, and the two answer keys are $1.00 each. Specimen sets for each test are $3.00. Date of information: 1978.

FIELD TEST DATA

Each test was field tested on over 800 ninth graders and 2400 twelfth graders in New Mexico. Difficulties and other statistics are reported for each item, as are norms and internal consistencies for the total test.

ADMINISTRATION

These are designed for group administration. Testing time is 20 minutes for each.

SCORING

Templates are available for hand scoring.

NEW MEXICO CONSUMER MATHEMATICS TEST & CONSUMER RIGHTS AND RESPONSIBILITIES TEST

Monitor, 1973

MEASUREMENT PROPERTIES

A B (C) 1. Description.

A (C) 2. Agreement.

A (C) 3. Representativeness.

A (C) 4. Sensitivity. An average superiority of about 2.5 items correct for 12th graders over 9th graders is reported, but that gain is not clearly attributable to instruction.

A (C) 5. Item Uniformity. Internal consistencies of .62 to .75 for the total tests are reported, but consistencies within content clusters of items are not given. Several items on each test (e.g., three for Consumer Rights and Responsibilities at grade 12, eight for Consumer Math at grade 9) have correlations with the total test score of less than .3.

A (C) 6. Divergent Validity.

A (C) 7. Bias.

A (C) 8. Consistency.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A B (C) 15. Flexibility.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. Templates for scoring are available.

A (C) 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data. Norms for pupils in New Mexico are given.

PRE-READING ASSESSMENT KIT — CTB/McGraw-Hill Ryerson Limited, 1972

DESCRIPTION

Pre-Reading Assessment Kit is designed as a "rough screening device for the classroom teacher" to use with children in kindergarten and first grade. Its tests measure skills in the following four areas: listening, symbol perception, experience vocabulary, and comprehension. The kit has tests at three levels of difficulty, the number of objectives ranging from three at the difficult level to eight at the easy one. Items are multiple choice, averaging ten per objective.

PRICES

A classroom set of consumable test forms for 32 pupils costs $1.67 per pupil and includes record forms and a manual. The manual is $2.40 separately. A specimen set is offered for $3.60. Date of information: 1977-78.

FIELD TEST DATA

Difficulty leveling was based on a pretest of 2864 first graders. It is likely that these pupils were Canadian.

ADMINISTRATION

These tests are made for group administration. Estimated time for each of the 18 subtests is 10 minutes.

SCORING

The manual contains keys for hand scoring.

COMMENTS

The manual suggests that tests like these are biased against children from limited English speaking or culturally disadvantaged background.

PRE-READING ASSESSMENT KIT

CTB/McGraw-Hill Ryerson Limited, 1972

MEASUREMENT PROPERTIES

A B (C) 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing. Resource materials are identified for some portions of the test, but the information is not detailed.

(A) B C 15. Flexibility. Each subtest is on a separate form.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Possible interpretations of scores are discussed and suggestions are given for cutting scores. Support for the decisions is not given.

A (C) 21. Comparative Data.

PRESCRIPTIVE READING INVENTORY CTB/McGraw-Hill, 1972

DESCRIPTION

The PRI is a six-level system for testing the following areas of reading skill: recognition of sound and symbol, phonic analysis, structural analysis, translation (meanings of words and phrases), literal comprehension, interpretive comprehension, and critical comprehension. Levels 1 and 2, for K to 1.0 and K.5 to 2.0 have 10 objectives each. The upper four levels, aimed at grades 1.5 through 6.5, have 34 to 42 objectives per level with an average of 3 to 4 multiple choice items per objective. In addition to the booklet for testing each level, smaller interim tests are optionally available for monitoring progress during the school year. The Interpretive Handbook (included) has guidelines for integrating the PRI into instruction and suggestions for classroom activities for each objective. Guides indexing the PRI to basal reading series are optionally available.

PRICES

Test booklets in sets of 35 sell for various prices depending on whether they are reusable (for the upper two levels, 39¢ to 44¢ each), hand scorable (57¢), or machine scorable (71¢). Answer sheets are 10¢ each in packs of 50. Keys for hand scoring are 16¢ per pupil in sets of 35. One per pupil is needed. Included in the specimen set ($5.50 for each level, $11.00 for all levels) are test booklets, answer sheets, plus the following materials, with their separate prices in parentheses: examiner's manual ($2.50 per level), and an Interpretive Handbook ($3.25). A Technical Report is available for each level at $3.25. Date of information: 1979.

FIELD TEST DATA

A national tryout was conducted on an ethnically mixed national sample of 18,000 students. In the Technical Report, several analyses of these data are presented, including a comparison of difficulties for "standard" and Black samples of pupils. Reliability, validity, and sensitivity to instruction data are given. Data are also given for the study equating the PRI and the CAT-70.

ADMINISTRATION

The PRI is a group test. Time for testing an entire level is about three hours. The publisher recommends administering the lower two levels by cassettes.

SCORING

The basic scoring service, which costs 70¢ per pupil for answer sheets or 97¢ per pupil for scoring booklets, reports individual scores and group summary scores by objective. Estimates of normative scores are optionally available. Estimated reporting time is 15 days from receipt by the publisher. Hand scoring keys are provided in the Interpretive Handbook for all levels.

PRESCRIPTIVE READING INVENTORY CTB/McGraw-Hill, 1972

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. Item sensitivity data provide a very rough indication of degree of agreement.

A (C) 3. Representativeness.

A (C) 4. Sensitivity. Average item sensitivities of .20 - .38 per level are reported for the tryout version of the PRI using the index of Marx and Noll. Data are not reported at the level of the item or objective.

(A) C 5. Item Uniformity. KR-20 reliability coefficients range from .63 to .88.

A (C) 6. Divergent Validity. Reported factor analyses do not support the separateness of the tested skills in a consistent fashion across test levels.

A (C) 7. Bias.

(A) (C) 8. Consistency. For the tests of 34 objectives, a type of alternate form reliability is reported, namely correlation of the scores for an objective with scores on a longer criterion test of the same objective. Seven to ten objectives from each level were sampled. Data are reported for each of 2-3 grades for each test level.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review. Item analysis and revision were done after tryout.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

(A) C 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. There is a good carryover of objectives across levels, but single objectives are not necessarily easy to test separately. Optional interim tests give more flexibility.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

(A) B C 18. Scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Three levels of attainment are identified, but with little justification.

(A) C 21. Comparative Data. Pupils' performance on the PRI may be used to estimate their performance on the California Achievement Test in normative terms, when the publisher's scoring service is used.

READING: COMPREHENSION SKILLS (K-6)

Instructional Objectives Exchange, 1973

DESCRIPTION

This battery is a collection of multiple choice tests measuring 40 objectives in reading comprehension. The objectives deal with the following groups of skills: main idea (10 objectives), conclusions (10), sequence (7), context clues (9), punctuation (3), syntactical structures (4), affixes (2). The five to ten items per objective are printed on spirit masters for local duplication and scoring. Two alternate forms of this collection are sold.

PRICES

Each form of this collection sells for $29.95 which includes a manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used. Date of information: 1979.

FIELD TEST DATA

A formative field test is mentioned but not described. After publication, performance data were gathered on 81 to 737 pupils per objective (average: over 500).

ADMINISTRATION

These are group tests.

SCORING

Answer keys are provided in the manual for hand scoring.

COMMENTS

Leveling of tests in this collection according to content, format, field test data, etc., is done locally by the test user. The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C 1. Description. Amplified objectives, but without rules for sampling the domains.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. The keys for all 40 objectives are printed on one page in small type.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A B (C) 21. Comparative Data. Comparative data are given in the form of cumulative percentages of pupils attaining each possible score for each objective at each of several separate grades. The sample is all from Southern California.

READING: WORD ATTACK SKILLS (K-6) Instructional Objectives Exchange, 1973

DESCRIPTION

This battery is a collection of multiple choice and oral response tests measuring 38 objectives in word attack. There are five to ten items per objective (mostly ten), items for each objective being printed on a separate spirit master for local duplication and scoring. Two alternate forms of this test are sold.

PRICES

Each form of this collection sells for $29.95. This price includes a manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used. Date of information: 1979.

FIELD TEST DATA

A small developmental field test is reported but not described. After publication, performance data were gathered on 81 to 713 pupils per objective (average: over 300).

ADMINISTRATION

These are group tests.

SCORING

Answer keys are provided in the manual for hand scoring.

COMMENTS

Leveling of the tests in this collection according to content, format, etc., is done locally by the test user. The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C 1. Description. Amplified objectives, but without rules for sampling the domain.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. The keys for all 38 objectives are printed on one page in small type.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data. Comparative data are given in the form of cumulative percentages of pupils attaining each possible score for each objective for three separate grades (on the average). The sample is all from Southern California.

REAL: READING/EVERYDAY ACTIVITIES IN LIFE — Cal Press, Inc., 1972

DESCRIPTION

REAL is a test of basic literacy skills for readers age 10 and above. It consists of 45 fill-in items, 5 each dealing with nine categories of common printed materials. For example, the category of "sets of directions" is tested by five items relating to a recipe for pizza which is given.

PRICES

Consumable test booklets are $1.00 each for orders of up to 100 copies. Cassette tapes for individual testing are $6.00 each. The Administrator's Manual, with technical information, is $6.50. A specimen set is available for $8.00. Date of information: 1977.

FIELD TEST DATA

After a developmental field test on 300 persons, mostly junior and senior high school students in inner city schools, REAL was revised and then normed on 434 disadvantaged Job Corps students of ages 18-21. Percentile norms, total test reliability (KR-20 = .93), point biserials for individual items, and item difficulties are given.

ADMINISTRATION

The REAL is administered to groups or individuals with the aid of cassette tapes and earphones.

SCORING

Scoring is done by hand using model answers in the manual.

REAL: READING/EVERYDAY ACTIVITIES IN LIFE — Cal Press, Inc., 1972

MEASUREMENT PROPERTIES

A B (C)	1.	Description.
A (C)	2.	Agreement. Content validation procedures are alluded to but not described.
A (C)	3.	Representativeness. An effort to ensure representativeness is alluded to but not described.
A (C)	4.	Sensitivity.
A (C)	5.	Item Uniformity. The internal consistency data are not at the level of the objective.
A (C)	6.	Divergent Validity.
A (C)	7.	Bias.
A (C)	8.	Consistency.

APPROPRIATENESS AND USABILITY

(A) B C	9.	Instructions.
(A) C	10.	Item Review.
(A) C	11.	Visibility.
(A) C	12.	Responding.
(A) C	13.	Informativeness.
A (C)	14.	Curriculum Cross-Referencing.
A B (C)	15.	Flexibility.
A (C)	16.	Alternate Forms.
(A) C	17.	Administration.
A B (C)	18.	Scoring.
(A) C	19.	Record Keeping.
A (C)	20.	Decision Rules.
A (C)	21.	Comparative Data. The norm sample is small.

SIPAY WORD ANALYSIS TESTS
by Edward R. Sipay

Educators Publishing Service, Inc., 1974

DESCRIPTION

The Sipay Word Analysis Tests consist of a 17-test diagnostic battery measuring word-analysis skills in these three broad areas: visual analysis, phonic analysis, and visual blending. The tests range in breadth from "visual analysis" with three subtests and a total of 99 items, to "vowel sounds of y" with 9 items. There are at least three items for each specific skill (e.g., contractions with *not*), the items all calling for oral responses. The first test is a 57-item diagnostic survey.

PRICES

This battery is sold for $73.00 in a kit which includes a manual, a "mini-manual" for each of the 17 tests, 12 answer sheets for each test, and a set of 756 stimulus cards. Answer sheets are available separately in sets of 12 for 15¢ to 60¢ depending on test length. Specimen sets are $2.50. Date of information: 1977.

ADMINISTRATION

The Sipay tests are made for administration to individuals by a teacher.

SCORING

The pupil's oral responses are scored by teacher judgment at the time of responding.

COMMENTS

The stimuli, when they are words or syllables, are chosen to be uncommon so that they are unlikely to be in children's sight vocabulary. The developer disagrees with our rating of feature #18 and says that many users do not find the directions for the examiner (feature #17) complicated.

SIPAY WORD ANALYSIS TESTS
by Edward R. Sipay

Educators Publishing Service, Inc., 1974

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

(A) C 3. Representativeness. Principles for selecting stimuli are described in detail. A number of the domains are tested in full, not merely sampled.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. Although there are no field test data, specific instructions are given to avoid scoring dialect responses as incorrect.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Sample items are not given for about half of the tests.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

A (C) 16. Alternate Forms.

A (C) 17. Administration. The directions for administering, scoring, and interpreting results are complicated.

A B (C) 18. Scoring. The recording and scoring of responses is often complex and subjective.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Cutoff points are given but without support.

A (C) 21. Comparative Data.

SKILLS MONITORING SYSTEM: READING — Harcourt Brace Jovanovich/ The Psychological Corporation, 1975

DESCRIPTION

The SMS: Reading is a four-level instructional management system for reading which measures pupils' skills in word identification at a Grade 3 level (including visual perception, phonics, morphemic elements) and comprehension at 3rd, 4th, and 5th grade levels (including word meaning in context, literal meaning, interpretation, critical reading). Each level includes both "locator" or diagnostic tests of from 27 to 36 objectives, with two multiple choice items per objective, and "skill-minis" for the same number of objectives with 8 to 12 items per objective. Practice skill-minis are also available.

PRICES

At each level, a package of 35 skill locators with scoring key, class record, and teacher handbook is 55¢ to 69¢ per pupil for the machine scored form. Keys for hand scoring are $1.35 per level. Self-scoring skill-minis are 17¢ per pupil in sets of 16. A classroom set of materials is also sold. Specimen sets are $2.75 per level. Date of information: 1978.

FIELD TEST DATA

Publisher reports that the SMS: Reading was field tested on roughly 6000 pupils in 215 classrooms at grades 3, 4, and 5 in selected school systems.

ADMINISTRATION

These are designed as group tests.

SCORING

Machine scoring of the locator tests costs 75¢ per pupil. The locators and skill-minis may be scored by hand from a key, or the skill-minis may be ordered in a self-scoring form.

COMMENTS

An optional Teacher's Resource Notebook was in preparation in 1977. This will contain guidelines for instruction and an index of curricular resources.

SKILLS MONITORING SYSTEM: READING

Harcourt Brace Jovanovich/
The Psychological Corporation, 1975

MEASUREMENT PROPERTIES

A (B) C 1. Description.

(A) C 2. Agreement. Judges sorted test items into homogeneous groups, wrote objectives for each group, then compared their objectives with the original ones. The level of detail in those objectives and the method of comparing objectives are not described.

A (C) 3. Representativeness.

A (C) 4. Sensitivity. No data.

(A) C 5. Item Uniformity. Median KR-20s and ranges of KR-20s are reported for each test length in each level. Medians are mostly .73 - .83.

(A) C 6. Divergent Validity. The evidence is not strong: low correlations mostly (<.4) among pairs of items measuring different objectives on the locator tests.

A (C) 7. Bias. No data, but a review for bias is mentioned.

(A) C 8. Consistency. A type of alternate form reliability is reported in very general terms: median tetrachoric correlations for each level between the mastery judgment on the locator for each objective and the corresponding judgment on the skill-mini. Values range from .67 to .73.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Sample items are generally not given.

(A) C 10. Item Review. Item selection and revision were based on field test data.

(A) C 11. Visibility.

(A) C 12. Responding. Also, latent image format of minis gives instant feedback.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing. In preparation.

(A) B C 15. Flexibility.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

(A) B C 18. Scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Decision rules are given but without support.

A (C) 21. Comparative Data.

SOCIAL STUDIES: AMERICAN GOVERNMENT (10-12)

Instructional Objectives Exchange, 1973

DESCRIPTION

This battery is a collection of multiple choice tests measuring 32 objectives in American government. An average of three to four of the objectives deal with each of the following topics: our colonial heritage, the Constitution, citizens' rights, politics, the Congress, the Executive, the Federal Judiciary, and state and local government. Each test item is printed on spirit masters for local duplication and scoring. Two alternate forms of this collection are available.

PRICES

Each form of this collection sells for $29.95. This price includes a manual and record forms. The price per pupil will vary with the number of copies made from each spirit master and the number of objectives that are used. Date of information: 1979.

FIELD TEST DATA

Field testing in one high school is mentioned but not described.

ADMINISTRATION

These are group tests for administration by a teacher or by oneself.

SCORING

Answer keys are provided in the manual for hand scoring.

COMMENTS

The publisher also offers a customized CRT service.

MEASUREMENT PROPERTIES

(A) B C 1. Description. Amplified objectives, but without rules for sampling the domain.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. No data.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

A B (C) 9. Instructions. The language of instructions and stems may be difficult for the average high school student.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. Answers to all 32 tests are printed on one sheet in small type.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data.

SRA SURVIVAL SKILLS IN READING AND MATH — Science Research Associates, 1976

DESCRIPTION

The SRA Survival Skills Test is a 120-item test of practical problems in reading and math for pupils at grade 6 and above. For each of the 20 objectives in reading and 20 in math, there are 3 multiple choice items.

PRICES

Reusable test booklets are 73¢ each in sets of 25 (55¢ to schools) and answer sheets are 13¢ each by the 100. The administrator's manual is 70¢. A review set is offered for $1.30. Date of information: 1977.

FIELD TEST DATA

A technical report is available from SRA giving item difficulties and item/test correlations. Data are reported for a median of 560 pupils per grade for grades 7-12.

ADMINISTRATION

This test may be administered to groups.

SCORING

Machine scoring is offered at a cost of 98¢ per pupil, which includes the cost of answer sheets.

SRA SURVIVAL SKILLS IN READING AND MATH

Science Research Associates, 1976

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

A (C) 5. Item Uniformity. Point biserial correlations for items have a median near .45 for reading and .5 for math as given in the technical report. These are correlations of item scores with total test scores, not with scores for each item's objectives.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. No data.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A B (C) 15. Flexibility.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

(A) B C 18. Scoring.

A (C) 19. Record Keeping. There is no form for recording scores by hand.

(A) C 20. Decision Rules. For each three-item objective, the probabilities of a pupil getting scores of 0-3 by guessing are given.

A (C) 21. Comparative Data. Item difficulties are given in a technical report, but the sample of pupils that was tested is not described.

STANFORD DIAGNOSTIC MATHEMATICS TEST
by Leslie S. Beatty, et al.

Harcourt Brace Jovanovich/
The Psychological Corporation, 1976

DESCRIPTION

The Stanford Diagnostic Mathematics Test is a four-level battery testing skills that are usually taught in grades 1 through 8. Each level consists of three tests, one each dealing with number system and numeration, computation, and applications (problem solving, applications, tables, and graphs). At each level there are 11 or 13 objectives, there being an average of 8 to 10 multiple choice items per objective. Alternate forms are available.

PRICES

Hand scorable test booklets are 43¢ per pupil in sets of 35, these being reusable at the upper three levels. Keys for scoring test booklets are $3.85 per level and for scoring answer sheets $1.40 per level. Machine scorable and hand scorable answer sheets are about 11¢ each in sets of 35. Practice tests for each level are also offered optionally. Administrators' manuals are $2.75 per level. A standard package containing materials for testing 35 students is sold. Specimen sets are available at $3.30 per level. Date of information: 1978.

FIELD TEST DATA

The Stanford Diagnostic Mathematics Test was field tested on a national sample of 23,000 students and normed on a stratified sample of 36,000 pupils in grades 2-12. Percentile ranks, stanines, and grade equivalent scores are given as well as item difficulties for pupils at several separate grade levels per test level.

SCORING

Tests may be machine scored or scored by hand with a template. The basic scoring service runs 80¢ to 85¢ per pupil for machine scoring. The publisher estimates a turnaround time for test results of 10 working days from receipt.

STANFORD DIAGNOSTIC MATHEMATICS TEST
by Leslie S. Beatty, et al.

Harcourt Brace Jovanovich/
The Psychological Corporation, 1976

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. A review is mentioned but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity.

A (C) 5. Item Uniformity. Internal consistencies are reported for whole subtests (≥30 items) but not for separate objectives.

A (C) 6. Divergent Validity. High subtest intercorrelations suggest that aptitude and achievement are not well separated.

A (C) 7. Bias. No data, but editing to eliminate bias in the development of the tests is reported.

A (C) 8. Consistency. Alternate form reliabilities for clusters of items representing two to three objectives are reported for pupils at two separate grades per test level. Median tetrachoric coefficient is above .8.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. There is ample carryover of objectives from level to level, but objectives for one level are all in one booklet.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

(A) B C 18. Scoring. Templates and machine scoring options are available.

(A) C 19. Record Keeping.

(A) C 20. Decision Rules. Passing scores were set after considering several factors (e.g., whether a skill is a basis for later skills), but the process of setting these scores is described in very general terms.

(A) C 21. Comparative Data. See note on Field Test Data on facing page.

STANFORD DIAGNOSTIC READING TEST
by B. Karlsen, R. Madden, & E. F. Gardner

Harcourt Brace Jovanovich/
The Psychological Corporation, 1976

DESCRIPTION

The Stanford Diagnostic Reading Test is a four-level battery of tests designed to span grades 1.5 to 12.0. The following skill areas are covered: auditory discrimination, phonetic analysis, structural analysis, auditory vocabulary, word meaning, word parts, word reading, comprehension, rate, fast reading, and scanning/skimming. There are 17-25 objectives at each level with generally 6-8 multiple choice items per objective (range: 4 to 42 items). Alternate test forms are available. Publisher states that the SDRT places more emphasis on low achievers than is customary by including more than the usual proportion of easy questions. Guidelines for using the results for instructional and administrative purposes are given in the teacher's manual. A handbook referencing the tested skills to a variety of reading series is offered.

PRICES

Consumable test booklets are 43¢ each in sets of 35 for the lower two levels. Reusable booklets for the third and fourth levels are 43¢ and 48¢ in sets of 35. Answer sheets vary from 14¢ (hand scorable) to 28¢ (machine scored) in sets of 35. Scoring keys range between $3.00 and $3.60 per level, while each level of the manual for giving and interpreting the SDRT is $2.75. A specimen set is available at $3.30 for each level. Date of information: 1978.

FIELD TEST DATA

This revision of the SDRT was field tested on 24,000 pupils in grades 2-9 in 1974 and normed on a stratified national sample of 30,000 students in 1975. Percentile ranks, stanines, and grade equivalent scores are given as well as item difficulties for pupils at several different grade levels per test level.

ADMINISTRATION

The SDRT is a group-administered test battery. The estimated testing time for an entire level runs from 100 to 145 minutes.

SCORING

Scoring by hand template, key, or machine is available. The publisher's machine service costs from 85¢ to 90¢ per pupil. Publisher estimates a turnaround time for test results of 10 working days from receipt.

STANFORD DIAGNOSTIC READING TEST
by B. Karlsen, R. Madden, & E. F. Gardner

Harcourt Brace Jovanovich/
The Psychological Corporation, 1976

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness.

A (C) 4. Sensitivity.

(A) C 5. Item Uniformity.

A (C) 6. Divergent Validity. The subtests show high intercorrelations, which suggests they all measure the same thing.

A (C) 7. Bias. Data are not given, but editing for bias during test development is reported.

(A) C 8. Consistency. Alternate form reliability is generally above +.8.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review. Items were selected on the basis of field test data.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

(A) C 14. Curriculum Cross-Referencing.

A (B) C 15. Flexibility. There is a good overlap of objectives across levels but items for many objectives are intermixed, not grouped separately.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

(A) B C 18. Scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. "Progress indicator cutoff scores" are provided, but they are justified in only general terms. The publisher encourages local discretion in setting cutoffs. Use of normative scores for grouping is also explained.

(A) C 21. Comparative Data. Based on the national norming sample, percentiles, stanines, grade equivalents, and scaled scores are given.

SURVEY OF READING SKILLS

Dallas Independent School District, 1973-74

DESCRIPTION

The Survey of Reading Skills is an eight-level battery of tests measuring objectives in the following categories: pre-reading skills, structural analysis, word meaning, and comprehension. A test booklet and examiner's manual are provided for each of levels K-6. Level S, for secondary students needing remedial instruction, has four test booklets. The number of objectives per level ranges from 40 at S to 15 for 6th, the average number of items per objective ranging from 4 to 7.

PRICES

The price for the Survey of Reading Skills is the current printing and postage costs. The test booklets are consumable. Date of information: 1977.

FIELD TEST DATA

The system has been field tested, but results are not provided with the test materials.

ADMINISTRATION

The Survey of Reading Skills is designed for group administration, except for a second form of the K-level test.

SCORING

The tests are hand scored from keys in each examiner's manual.

COMMENTS

The difficulties of the levels are indicated by reference to specific texts in basal reading series. For example, Level II is aimed at the reading level of *Secrets* and *Rewards*; Level V at *Images*. The objectives themselves are commonly taught, not peculiar to this district.

MEASUREMENT PROPERTIES

A (B) C 1. Description.

A (C) 2. Agreement.

A (C) 3. Representativeness.

A (C) 4. Sensitivity.

A (C) 5. Item Uniformity.

A (C) 6. Divergent Validity.

A (C) 7. Bias.

A (C) 8. Consistency.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

A (C) 10. Item Review.

A (C) 11. Visibility. Graphics are often unclear.

(A) C 12. Responding.

A (C) 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing.

A (C) C 15. Flexibility. There is ample carryover of objectives across levels, but they are all tested in one booklet at each level.

A (C) 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. Hand scoring involves a complex chart/counting system.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Rules are provided without support.

A (C) 21. Comparative Data.

TESTS OF ACHIEVEMENT IN BASIC SKILLS - MATH	Educational and Industrial Testing Service, 1970-74

DESCRIPTION

TABS-Math is a seven-level battery of survey tests for pupils in grades K-12. There is one item per objective on all of the tests, objectives being grouped into the following clusters: arithmetic skills, geometry-measurement-application, and modern concepts. The number of items varies from 18 at Level K to 69 at the level for grades 4-6. The number of clusters per level is 3 or 2. Item formats are fill-in for levels K, 1, and 2, and multiple choice for the upper four levels. Alternate forms of this battery are available.

PRICES

Consumable test booklets for the lower four levels are 25¢ each in sets of 30. For the upper three levels, reusable booklets are 21¢ each in sets of 35 and answer sheets are 8¢ each in like sets. For each level the administrator's manual and answer key are each $1.50. A specimen set is offered at $2.25 per level. Date of information: 1977.

FIELD TEST DATA

The three test levels for grades 4-6, 7-9, and 10-12 were given preliminary tryouts and then were normed on national samples of 4500, 17,000, and 3500 pupils respectively. Means and standard deviations are reported for total test scores for four ability groups and three grade levels for each of those test levels. In addition, entry level item difficulties are given for all items at three grade levels.

ADMINISTRATION

The TABS are designed for group administration by a teacher.

SCORING

Hand scoring of the lower three levels is done with reduced pupil pages. Template and machine scoring are both offered for the upper three levels. The basic scoring service which costs 35¢ per pupil includes item and total scores for individuals and for classes.

TESTS OF ACHIEVEMENT IN BASIC SKILLS - MATH

Educational and Industrial Testing Service, 1970-74

MEASUREMENT PROPERTIES

A B (C) 1. Description. Objectives for single test items defeat the purpose of objectives, to describe skills and not single questions. The higher level clusters of items are described by extremely vague labels.

A (C) 2. Agreement.

A (C) 3. Representativeness.

A (C) 4. Sensitivity. Reported gains from grade to grade are not clearly the result of relevant instruction.

A (C) 5. Item Uniformity. Data reported are not for objectives or skill clusters.

A (C) 6. Divergent Validity.

A (C) 7. Bias.

A (C) 8. Consistency. At the level of the total test score, alternate form reliabilities are reported for two levels.

APPROPRIATENESS AND USABILITY

A B (C) 9. Instructions. Sample items are not provided, and the instructions for the lower levels are often unclear.

(A) (C) 10. Item Review. Quality control reported for the upper three levels only.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

A (C) 14. Curriculum Cross-Referencing. TABS is indexed to a curricular series of the publisher.

A (B) C 15. Flexibility. Good carryover of objectives across levels, but all are tested on one form.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

(A) B C 18. Scoring. Except at the lower three levels where the hand scoring materials are reduced pupil pages.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

(A) (C) 21. Comparative Data. For the upper three levels, there are detailed comparative data; for the lower four levels, none.

TESTS OF ACHIEVEMENT IN BASIC SKILLS - READING AND LANGUAGE	Educational and Industrial Testing Service, 1975

DESCRIPTION

The TABS is a three-level battery for assessing pre-reading and reading skills in pupils in grades K-2. There are 38 to 52 objectives per level which deal with the following categories of skill: word analysis, language development, comprehension, and study skills. A few affective objectives are included as well. For each objective there are from 1 to 24 items, the average being close to 3. Item formats include multiple choice, matching, and fill-in. A diagnostic and instructional program is available optionally. Two parallel forms of TABS are sold.

PRICES

Consumable test booklets with answer sheets are available for one test form at one level at 26¢ per pupil in a set of 30. The manual and answer key for a level are $1.50 together. For any one level the specimen set, test booklet plus manual, is $4.50. Classroom sets of the teaching and testing materials are available on approval. Date of information: 1977.

ADMINISTRATION

TABS are designed for group administration.

SCORING

Answer keys for hand scoring are available.

TESTS OF ACHIEVEMENT IN BASIC SKILLS - READING AND LANGUAGE

Educational and Industrial Testing Service, 1975

MEASUREMENT PROPERTIES

A (B) C 1. Description. Although written in the form of behavioral objectives, many of the objectives are vague.

A (C) 2. Agreement.

A (C) 3. Representativeness.

A (C) 4. Sensitivity.

A (C) 5. Item Uniformity.

A (C) 6. Divergent Validity.

A (C) 7. Bias.

A (C) 8. Consistency.

APPROPRIATENESS AND USABILITY

A (B) C 9. Instructions. Sample items are not provided.

A (C) 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

A (C) 13. Informativeness. Contents of the specimen set are not listed in the catalog. It is not clear which manuals are available.

A (C) 14. Curriculum Cross-Referencing. Keyed to the publisher's own instructional program.

A B (C) 15. Flexibility.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A B (C) 18. Scoring. Answer keys are not consistently easy to use. Some subjective judgments are involved in scoring.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Rules are provided without support. Some decisions are based on one item.

A (C) 21. Comparative Data.

WISCONSIN DESIGN FOR READING SKILL DEVELOPMENT: COMPREHENSION
by Wayne Otto, Karlyn Kamm, et al.

NCS Educational Systems, 1977

DESCRIPTION

The Wisconsin Design is a seven-level battery of measures for diagnosing the status and monitoring the progress in reading comprehension of pupils in grades K through 6. The number of objectives per level ranges from 3 to 8, with at least 12 items per objective. Thirty-three of the objectives in the battery have multiple choice items; six ask for written responses; one asks for oral responses. Fifteen different types of literal and interpretive comprehension are tested in all. Alternate forms are available. Optional supporting materials include a teacher's planning guide and teacher's resource file. This battery is one part of a six-part instructional management system; the word attack and study skills tests are also reviewed in this volume.

PRICES

Consumable test booklets for the lower grades are 59¢ to 80¢ per pupil and reusable booklets for the upper levels are $1.71, both types coming in sets of 35 along with an administrator's manual. The tests for the lower levels are also available on spirit masters at $16.00 to $27.00 per level. Spirit masters for printing answer sheets are $3.00 each. Specimen sets are $6.00. The teacher's planning guide is $4.25 and the teacher's resource file is $41.50. Date of information: 1978.

FIELD TEST DATA

Each multiple choice objective was field tested on about 150 pupils fairly evenly drawn from schools labeled low average, average, or high average in reading comprehension. Constructed response items were field tested on 8 to 24 pupils.

ADMINISTRATION

These tests are made to be given in groups by a teacher. Although the tests are not timed, the estimated time for testing a single skill is about 10 minutes.

SCORING

Keys are provided for hand scoring of multiple choice items. Models of correct responses are given for the constructed response items.

COMMENTS

Data for test features #1, 2, 4, 5, and 8 were provided by the publisher after the original test review was completed. The ratings here for those features were made by one person (CBW). The technical reports cited are available from the University of Wisconsin R&D Center for Cognitive Learning.

WISCONSIN DESIGN FOR READING SKILL DEVELOPMENT: COMPREHENSION
by Wayne Otto, Karlyn Kamm, et al.

NCS Educational Systems, 1977

MEASUREMENT PROPERTIES

(A) B C 1. Description. Given in Working Paper #213, a preview of the final technical manual.

A (C) 2. Agreement. A review for agreement is mentioned in Working Paper #213, but not described.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. Gains are reported in a paper by Karlyn Kamm, but spurious sources of increase are not clearly controlled.

A (C) 5. Item Uniformity. Publisher expected to have data available by the time this volume is published.

A (C) 6. Divergent Validity. No data.

A (C) 7. Bias. No data.

A (C) 8. Consistency. Publisher expected to have data available by the time this volume is published.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

(A) C 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. Hand scoring only.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules. Three levels of attainment are distinguished, but their rationale is not given in the test package. Publisher says that a dissertation by Demos deals with this issue.

A (C) 21. Comparative Data.

WISCONSIN DESIGN FOR READING SKILL DEVELOPMENT: STUDY SKILLS
by Wayne Otto, Karlyn Kamm, et al.

NCS Educational Systems, 1973

DESCRIPTION

The Study Skills component of the Wisconsin Design is a seven-level battery of tests for pupils in grades K through 6. The major content strands deal with pictures and maps, graphs and tables, and reference materials. There are from 2 to 14 objectives per level, each with at least ten multiple choice items per objective. Alternate forms are available for most of the tests in this battery. This battery is one part of a six part instructional management system; the comprehension and word attack tests are reviewed in this volume. Optional supporting materials include a teacher's planning guide and teacher's resource file.

PRICES

Consumable test booklets are from 28¢ to 80¢ per pupil for the lower four levels and reusable booklets are $1.71 for the upper levels, both types coming in sets of 35 with an administrator's manual. Tests for the lower levels are also available on spirit masters at $6.00 to $28.00 per set, depending on the number of separate tests that make up the level. Machine scorable answer sheets for the upper levels are printed locally from spirt masters which cost $3.00; the teacher's planning guide is $4.25; and the resource file plus supplement is $61.00. Date of information: 1978.

FIELD TEST DATA

After pilot-testing the precommercial edition in 22 schools and revising it, publisher field tested this edition in three schools of average achievement level in Georgia. Over 1000 pupils provided data, 455 taking alternate forms of a subset of objectives, and 605 taking adjacent levels of the battery. A variety of data are given including, for each objective, average correct, frequency distributions, and internal consistencies. Alternate form reliabilities and inter-level correlations are reported in several ways. Data appear in Working Papers #190, #391, and #422 which are available from the University of Wisconsin R&D Center for Cognitive Learning.

ADMINISTRATION

The Wisconsin Design Study Skills tests are made for group administration. Working paper #190 gives 14 minutes as the approximate average time for administering these untimed tests.

SCORING

Scoring is by hand key.

COMMENTS

Data for test features #1, 4, 5, 6, 8, and 21 were provided by the publisher after the original test review was completed. The judgments reported here for those features were made by one person (CBW).

WISCONSIN DESIGN FOR READING SKILL DEVELOPMENT: STUDY SKILLS
by Wayne Otto, Karlyn Kamm, et al.

NCS Educational Systems, 1973

MEASUREMENT PROPERTIES

(A) B C 1. Description. Given in Working Paper #190.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. Technical Reports #341 and #422 show gains in scores and levels, but spurious sources of increase are not clearly controlled.

(A) C 5. Item Uniformity. Median internal consistency (Hoyt r) per objective per level is close to .74 for form P.

(A) C 6. Divergent Validity. Intercorrelations of scores on pairs of objectives within a level are generally below .5. Intercorrelations of mastery decisions for all pairs of tests within levels and between adjacent levels are also given.

A (C) 7. Bias. No data.

A (C) 8. Consistency. Alternate form consistencies are given for only 24 objectives from the upper five test levels. These are in two forms: consistency of mastery decisions and of number correct. The median of the alternate form raw score correlations is r=.51 for these objectives.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

(A) C 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

(A) C 16. Alternate Forms. Available for most of the tests.

(A) C 17. Administration.

A (B) C 18. Scoring. Hand scoring only.

(A) C 19. Record Keeping. Class record sheets may need to be made locally, but individual records are provided.

A (C) 20. Decision Rules. A mastery percentage is given, but not supported.

A (C) 21. Comparative Data. Although a variety of data are given, Working Paper #190 says that they are not intended for use as norms. The sample of pupils is geographically limited.

WISCONSIN DESIGN FOR READING SKILL DEVELOPMENT: WORD ATTACK
by Wayne Otto, Karlyn Kamm, et al.

NCS Educational Systems, 1972

DESCRIPTION

The Wisconsin Design Tests of word attack are a four-level battery for diagnosing the status and monitoring the progress of pupils in grades K-6. The objectives deal mainly with readiness, phonics, sight reading, and structural analysis. There are from six to sixteen objectives per level, each having at least fifteen multiple choice items. Alternate forms are available. Optional supporting materials include a teacher's planning guide and a teacher's resource file. This battery is one part of a six part instructional management system, the comprehension and study skills tests are also reviewed in this volume.

PRICES

Consumable test booklets are 29¢ to 59¢ per pupil in either a self-scoring or hand-scoring format. Tests are also available on spirit masters at $10.50 to $18.00 per level. The examiner's manual which is included with each set of consumable test booklets, is available separately for $1.60 per level; teacher's planning guide is $4.25; and the resource file plus supplement is $61.00. Date of information: 1978.

FIELD TEST DATA

After pilot testing this battery in 23 schools and revising it, publisher field tested this version in three schools in New York. A median of 152 pupils per test level provided data for both test forms, and a total of 113 pupils were tested on two adjacent levels of the battery. A variety of data are given including, for each objective, average correct, frequency distributions, and internal consistencies. Alternate form reliabilities and inter-level correlations are reported in several ways. Data appear in Working Paper #190 which is available from the University of Wisconsin R&D Center for Cognitive Learning.

ADMINISTRATION

Thirty nine of the 45 objectives in this battery are designed for group testing by a teacher. Although the tests are untimed, the estimated time for testing a single skill averages 12 minutes (Working Paper #190).

SCORING

Scoring keys are provided.

COMMENTS

Data for test features #1, 5, 6, 8, and 21 were provided by the publisher after the original test review was completed. The judgments reported here for those features were made by one person (CBW).

WISCONSIN DESIGN FOR READING SKILL DEVELOPMENT: WORD ATTACK
by Wayne Otto, Karlyn Kamm, et al.

NCS Educational Systems, 1972

MEASUREMENT PROPERTIES

(A) B C 1. Description. Given in Working Paper #190.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. No data.

A (C) 4. Sensitivity. No data.

(A) C 5. Item Uniformity. Internal consistencies for individual objectives have a median of about .77 (Hoyt r).

(A) C 6. Divergent Validity. Intercorrelations of raw scores for pairs of objectives within each level are mostly below .5. Intercorrelations of mastery decisions for all pairs of tests within levels and between adjacent levels are also given.

A (C) 7. Bias. No data.

A (C) 8. Consistency. Median correlation between raw scores on alternate forms of single objectives is .64. Data are also given for consistency of mastery decisions across both forms of all objectives.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness.

(A) C 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility. Tests are available on spirit masters for separate duplication if desired.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. Hand scoring only.

(A) C 19. Record Keeping.

A (C) 20. Decision Rules.

A (C) 21. Comparative Data. Although a variety of data are given, Working Paper #190 says that they are not intended for use as norms. The sample of pupils is geographically limited.

WOODCOCK READING MASTERY TESTS
by Richard W. Woodcock

American Guidance Services, Inc., 1973

DESCRIPTION

The Woodcock Test consists of 400 oral response items for measuring the following reading skills in grades K-12: letter identification (45 items), word identification (150), word attack (50), word comprehension (70 analogy items), and text comprehension (85 modified cloze items). In each skill area, items are arranged in ascending difficulty as determined by Rasch-Wright item analysis methods. Pupils work the test from their own basal level to their own ceiling. Alternate forms of the Woodcock are available.

PRICES

A complete set of materials for either form of the test costs $22.00. It includes the easel kit with all of the test items, the manual, and 25 forms for scoring and interpreting responses. Date of information: 1978.

FIELD TEST DATA

The final pool of 800 items (400 per form) was selected from an initial pool of over 2400 as a result of developmental testing. The final tests were normed on a fairly representative national sample of over 5000 pupils.

ADMINISTRATION

The Woodcock is an individual test which can be administered by a classroom teacher in an estimated time of 20 to 30 minutes.

SCORING

Individual responses are scored and recorded on the spot as the student speaks them. Correct answers are visible to the examiner on the backs of easel kit stimulus cards.

COMMENTS

Fall and spring percentile norms and normal curve equivalents for the Woodcock tests were expected to be available by the time this volume is published.

WOODCOCK READING MASTERY TESTS
by Richard W. Woodcock

American Guidance Services, Inc., 1973

MEASUREMENT PROPERTIES

A (B) C 1. Description. Although not stated in the usual form of behavioral objectives, the domains are described fairly well in the manual.

A (C) 2. Agreement. No data.

A (C) 3. Representativeness. Items were selected on statistical grounds.

A (C) 4. Sensitivity. No data.

(A) C 5. Item Uniformity. Split half reliabilities for 103 pupils on the five subtests vary from .79 to .99 at grade level "2.9." On the four tests of word- or text-level skills, they range from .83 to .98 at the "7.9" grade level for 102 pupils.

A (C) 6. Divergent Validity. Tables 10-14 in the manual report correlations between subtests and of subtests with the total for other tests in the battery. They are rather dependent at the lower grade levels, over half the correlations being $\geq$ .7. At the upper grade levels, relative independence is shown.

A (C) 7. Bias. No evidence.

(A) C 8. Consistency. Reliabilities for retesting with the alternate form are .84 or better at the subtest-level in 7 out of 10 cases reported.

APPROPRIATENESS AND USABILITY

(A) B C 9. Instructions.

(A) C 10. Item Review.

(A) C 11. Visibility.

(A) C 12. Responding.

(A) C 13. Informativeness. Specimen sets are not offered, but the materials may be returned for refund within 30 days if they are in unused condition.

A (C) 14. Curriculum Cross-Referencing.

(A) B C 15. Flexibility.

(A) C 16. Alternate Forms.

(A) C 17. Administration.

A (B) C 18. Scoring. Scoring of item responses is generally easy and objective, but may require some judgments of meaning. Converting the raw scores to derived scores requires some practice.

(A) C 19. Record Keeping.

(A) C 20. Decision Rules. The decision rules are like confidence intervals and predictions of success in using material at specific levels of difficulty.

(A) C 21. Comparative Data.

CHAPTER 5
How To Select Tests: Locating Tests and Comparing Their Technical and Practical Features

This is the first of two chapters on selecting a test so that it will be suited to the needs of a particular program. This chapter describes procedures for locating and screening tests to arrive at a number that is workable for evaluating in detail. Methods for evaluating tests' technical and practical features, and comparing them according to these features are then given. A major concern in test selection--finding the one which best matches a specific curriculum--is covered in detail in Chapter 6.

Ideally a test user would be able to identify the single best test for a given need (for example, diagnosis of word attack skills of third graders in the inner city) by consulting a reference book of test evaluations. A number of factors make this method unfeasible. For one, ongoing developments in testing cause a reference work to grow obsolete starting at the time when the research for the book stops. Second, not all features of a test are equally important to all test users, and a single test seldom excels in all features. Thus, it is necessary for individual users to weigh the various features according to their own needs and then to make overall comparisons. Finally, the single most important aspect of a test--its relevance to the test user's curriculum--can only be judged locally, by the people most familiar with that curriculum.

Before selecting a test, local program staff should decide whether testing is, in fact, the most effective response to their needs for information. This decision will depend on such issues as these:

- What type of information is needed?
- Who will receive the information?
- What other methods are there to obtain the information?
- What dollar costs are acceptable if the staff decides to proceed with a testing program?
- What costs in pupil and staff time are acceptable?
- How useful (i.e., timely and relevant) will the test scores be for the classroom teacher?

If testing turns out to be the preferred action, and if a specific test is not mandated by external authority, then test selection can proceed.

Having decided to test, most schools or districts seek to purchase ready-made tests--a logical first step. If suitable tests are not available, however, two other options can be considered. First, tests or testing systems may be created locally. The considerable cost in staff time for such a project may be substantially reduced through the use of such resources as skills continuums, objectives collections, and item banks (see Appendix A). The benefits of maintaining local control over testing may offset the costs of this option. Because the development of a test battery is a long-range project, this option should be followed only after careful consideration of the alternatives.

A second option is to hire a test developer to create a testing system. A number of publishers will custom-make CRTs. Appendix A lists some of these publishers. The tests they produce should be subject to the same evaluation procedures that would be applied to ready-made tests under consideration.

The procedures described in this chapter are meant to help you assess the merits of available NRTs, CRTs, or a mixture of the two, in your search for a test. Although one could use these procedures to examine a single test or test package, they are most useful for comparing two or more tests. It is not possible to say how much overall quality a single test must have in order to be "good enough," nor is it possible to determine that the match of a single test to a given curriculum is "close enough." One can only decide which of several tests is better.

Many potential test buyers will not have the personnel to follow all of the procedures in this chapter and the next one. We have included them so that test users can make decisions consciously rather than by oversight. Where test selection is carried out by a committee, as it is in a majority of school districts in the United States,[1] it will be easier to evaluate tests thoroughly before choosing one.

Test selection involves a number of technical decisions, so it is *essential*[2] that some of the people involved in the process have a knowledge of the principles of both criterion-referenced and norm-referenced testing. To maximize the instructional relevance of testing and to minimize the possible alienation resulting from it, it is also important to involve teachers and curriculum specialists--those most familiar with the students and the curriculum--at every step of test selection.

Finally, a word should be said about the importance of local field testing in test selection. Though not always possible, it is extremely helpful to try out a test in your own schools before deciding to adopt it. Teachers' and pupils' reactions to a test are very significant indicators not only of its appropriateness for your setting, but also of its quality and usability. Local test tryouts may serve either to screen out less desirable measures or to choose one out of a pool of finalists in the selection process.

---

[1]Dotseth, et al., 1978.

[2]APA, 1974.

IDENTIFY TESTS WHICH SEEM APPROPRIATE AND DO AN INITIAL SCREENING

Before starting to search for tests, you should be clear about your purposes in testing. For some purposes, certain characteristics of a test will be more important than others. A good understanding of what kind of information you want from the test will help you identify the test characteristics which are most important for your purposes.

Any purpose for testing is best described in terms of a type of decision which the test results are meant to influence. For example, a common purpose is to select a limited number of individuals from a large pool of available students, as in selecting for admission to a special program. Another purpose is to guide the planning of instruction by measuring students' current proficiency on a given set of skills. Still another is to make decisions about individual students by measuring how well they have mastered the objectives of a program.

Once the purpose for testing is made clear, you can develop a pool of available tests by means of a systematic search process. A good starting point for the search is the set of test reviews in Chapter 4 and in the reference works listed in Appendix B. Information in any reviews may need to be updated by referring to test publishers' current catalogs which are readily obtainable by mail.

At this point in the test selection process, you are working from *descriptions* of tests. As you look through these materials to exclude tests which do not respond to your specific needs, you are doing an initial sifting to arrive at a manageable number of tests for closer consideration.

To help you with this initial sifting, the following paragraphs mention several test uses and their implications for test selection.

Testing for diagnosis and prescription of the individual student

In order to be most usable for diagnosing individuals' strengths and needs, and for assigning lessons, a test must have these qualities:

- Test items are keyed to clear and teachable objectives.
- There are several items per objective.
- Hand scoring is practical for quick use of results.
- A score is given for each objective.
- If scoring is by machine, the return of results to teachers is rapid, and score reports are easy to interpret.

Tests with only two or three items per objective will save testing time, but their consistency in identifying individual students' strengths and needs on particular objectives is lower than that of tests with more items per skill. Diagnostic tests may be packaged to allow testing only a small number of objectives at once, but usually they survey a large number of objectives in one test booklet.

It is up to the test buyer to decide what level of subject matter detail is needed in the test scores to support diagnosis and prescription.

Some educators believe that scores on fairly broad content areas such as *vocabulary*, *word attack*, and *critical thinking* are useful. Most classroom teachers feel that scores for objectives are needed at the level of a lesson or small number of lessons.

Testing to verify or monitor ongoing student progress

The traditional tool for monitoring students' learning is the teacher-made test. On the basis of the test scores, students are moved forward to the next lesson or are given more practice on the current one. A number of test publishers have produced batteries of many short tests which are meant for the same purpose. To be well suited for this purpose, a test battery must have these qualities:

- Test items are keyed to clear and teachable objectives.
- The test is packaged to allow testing a small number of objectives at one sitting, preferably one objective.
- There is an adequate number of items per objective.
- Hand scoring is practical for quick use of results.
- A score is given for each objective.

These tests differ from diagnostic ones by covering a very small number of objectives in each test form to permit flexible, individualized testing of specific lessons as they are taught. Verification of student progress also requires a very reliable score on each skill so as to be sure of each student's degree of learning; a reliable score, in turn, requires fairly large numbers of items per objective.

Many instructional programs in reading and math have progress-monitoring test batteries as optional components. These batteries need to be evaluated before purchase just as carefully as any other tests.

Testing for program planning or needs assessment

When testing is done to identify the strengths and needs of a given curriculum, it can be thought of as diagnostic testing at the program level. Such tests should:

- Survey the appropriate range of content and skills.
- Give scores that allow planning decisions to be made.

Breadth of coverage is relevant here, not reliability at the level of the individual student. Thus the number of items per objective that individuals answer need not be large. Presumably, scores on tests for diagnosing individuals could be aggregated and used for this planning function, allowing the test to serve two purposes at once.

Testing for program evaluation or accountability

When testing is conducted to meet external requirements, those requirements may state which characteristics the test should have or even which test to use. Any required characteristics, such as the presence of national norms or of other field test data, can be used as screens in test selection. A growing number of CRTs provide norms along with the absolute scores. Testing for the purpose of program evaluation usually calls for the use of measures which survey a broad range of content and skills. If the choice of test is left to local discretion, then the test should also

give scores that will support instructional decision making, at least at the program level if not at the classroom level. If instructional relevance is a lost cause, then tests for accountability or program evaluation can be chosen so as to minimize testing time.

Testing for other purposes

A few other obvious or surface features can be used for eliminating tests at this preliminary stage when you are working from the test descriptions--for example, availability of alternate test forms (for pre- and posttesting purposes). There will not be enough information in secondary sources to inform many other test selection decisions, although it may seem that there is. Take, for example, the need to select students for a special program. If an NRT is to be used, the appropriateness of the norm group is crucial. But information on norm groups is not available in many test reviews nor in most publishers' catalogs. In most cases, the test package itself will have to be examined directly in order to make judgments about other critical test features.

EXAMINE SPECIMEN SETS

Once some promising tests have been found in the secondary sources, specimen sets of these tests should be ordered. Further selection is then done by referring to actual test materials and manuals. At least two broad standards should be applied at this point.

Standard 1

First, the cultural appropriateness of each test's items for your student population should be judged. Some of the questions that will help you gauge the appropriateness of a test's items are these:

- Are the concepts familiar to your students?
- Is the dialect of the language familiar to your students?
- Is the test's content free from social stereotypes?
- Are the instructions to the student understandable?

This standard can be applied effectively by classroom teachers and curriculum coordinators who have a good sense of what is culturally suitable for the program's students.

Standard 2

The other standard is a rough measure of a test's relevance to the local curriculum. Because the *objectives* of most existing tests--CRTs and well as NRTs--are stated rather loosely, they may seem to fit any curriculum. In order to judge how well the test materials cover the skills of your specific program, you should examine the actual test *items*.

For each of the tests under consideration, identify and count the items which measure skills that are actually taught in your program at roughly the same level. Record that number and then calculate the proportion of items on each test that are relevant to your program. Compare the tests on these two figures --total number, and proportion of locally relevant items. Eliminate the tests which have markedly lower figures. This task can be effectively carried out only by persons who are very familiar with the curriculum as it is actually taught.

This initial method of comparing tests with the local curriculum, while useful, is not adequate for finding the one test which is *best* matched to your program, for the following reasons:

- By matching test materials with the curriculum "as you remember it," you may overlook which and how many objectives of your own program each test *fails* to cover. In other words, since the focus is on test materials, skills in your program that are missing in the test battery will tend to be overlooked.

- By making global judgments of the relevance of test items, you may not attend to a number of other factors that affect the appropriateness of the test materials, such as difficulty of test items, appropriateness of item formats, and the relative importance of the skills which each test covers and does not cover.

This initial method for judging tests' curricular relevance is only a broad screening device. In Chapter 6, a more thorough method is given which takes into account the other factors that were just noted.

As mentioned earlier, steps up to this point in test selection should quickly reduce the tests under consideration to a number that is practical to evaluate in detail. If the number of tests remaining at this point is too large for available staff to study closely, then other test features may be used as screens, or the previous features may be reapplied more stringently. On the other hand, if the remaining pool of tests provides no satisfying choices, then serious thought should be given to developing tests locally, modifying existing tests, or not testing at all.

COMPARE TESTS ACCORDING TO THEIR PRACTICAL AND TECHNICAL MERITS

The method for comparing tests' merits that is outlined here involves selecting test features to evaluate, making judgments about those features, converting the judgments into numbers, combining the numbers for each test, and comparing tests in terms of the numerical totals. These steps may at first seem too detailed and mechanical. Three points should be noted in this regard.

First, by assigning numbers at each stage of judgment and carrying them to the next stage, you ensure that information from earlier judgments is not lost. In other words, the component decisions all have an effect on the final ratings of each test. Second, the methods are explicit. Therefore, they are teachable, repeatable, and easy to adapt. Finally, as you follow the steps, you will find the procedures are harder to read about in the abstract than they are to apply in a practical situation and that they become quite easy with a little practice.

As a practical matter, it is desirable to have specific features of tests evaluated by staff members who have the special training and experience to evaluate them. Thus your specialists in testing could evaluate tests' statistical qualities while teachers and curriculum specialists could judge the features, such as directions to the pupils and quality of prescriptive aids, which require a knowledge of pupils and instructional materials.

Table 2 summarizes the steps in comparing tests feature by feature and serves as a checklist for carrying out these steps.

TABLE 2
Checklist of Steps for Comparing the
Technical and Practical Merits of Tests

	Step		Page
____	1	Select test features to evaluate.	162
____	2	Rate the importance of the test features, and record the ratings on the *Worksheet*.*	162
____	3	Write the names of the tests to be compared at the top of the *Worksheet*, and duplicate the form for the test rater(s).	163
____	4	Find, in the sample materials for each test, the evidence for the first test feature.	168
____	5	Arrange the tests in descending order of merit on the given feature. Record these rankings (best, second, third...) in the respective columns of the *Worksheet* next to the name of the feature.	168
____	6	For tests which are equally good on a feature, give them the average of the ranks they would have earned if not equal. For tests which differ, but not by much, use the given rules of thumb.	168
____	7	Repeat Steps 4-6 for all other test features to be evaluated.	169
____	8	Summarize the tests' rankings by weighting them and then recording them in the "Final Results Table" at the upper right of the *Worksheet*.	169
____	9	Check that the total number of tallies per test in the "Final Results Table" is equal.	170
____	10	Compare the tests' profiles in the "Final Results Table." Eliminate tests that are markedly worse. Select the better ones for detailed analysis of their congruence with the local curriculum.	170

*Figure 1, pages 164-165.

STEP 1. Select test features to evaluate.

Two lists of test features for comparing tests have been developed at the Center for the Study of Evaluation--one for use with norm-referenced tests,[3] and the other for use with CRTs. The latter--the one used for test evaluations in this volume--is shown in Table 1, pages 14-15. Another list for evaluating CRTs was developed by Hambleton and Eignor.[4]

Any ready-made list should be edited by local staff. Such editing requires that the list be reviewed to determine if there are features you wish to add or omit. Features that should be omitted are:

- Ones that do not make a test better or worse for meeting your testing needs. These are features which are irrelevant or are of negligible importance. For example, the two test features, *curriculum cross-referencing* and *alternate forms*, may be eliminated from the judging process when there is to be one-time survey testing for accountability purposes, with its broad normative scores and slow reporting of results.

- Features that have already been used in a pass/fail fashion to narrow the pool of available tests. These are called *exclusionary features*. In screening tests to use for diagnosis, for example, you will already have excluded tests which do not provide *scores for separate objectives*.

Some features may be used in both a pass/fail fashion and a comparative one. For example, tests with fewer than some minimum acceptable number of items per objective may be excluded in the earlier screening; then, when tests are compared feature by feature, tests with larger numbers of items per objective may be rated higher than tests with smaller numbers. In the same vein, tests which do not offer optional *curriculum indexes* may be screened out, and the remaining tests later compared on the quality of their curriculum indexes.

Figure 1, pages 164-165, is a worksheet for recording and summarizing judgments about individual test features. In the first column of the worksheet, write the names of the features to be evaluated. Figure 2, pages 166-167, shows a worked example of the worksheet with a small set of features chosen from Table 1, pages 14-15.

STEP 2. Rate the importance of the test feature, and record the ratings on the worksheet.

A test's suitability for your needs depends more heavily on some of its features than on others. Three degrees of importance in features have already been recognized up to now:

- Exclusionary features--ones that are so important that they are essential if a test is to meet your needs. These are used in a pass/fail fashion to exclude clearly unacceptable or irrelevant tests.

- Irrelevant or unimportant features --ones that have just been eliminated from consideration because they do not make a test better or worse for your purposes.

[3]Hoepfner, et al., 1976.

[4]Hambleton and Eignor, 1978.

- Comparative features--all of those aspects of a test which make it more or less suitable. These include exclusionary features on which tests may still vary in quality even after they have met minimum levels of acceptability as mentioned under Step 1. Also included, of course, are various other test features you have deemed useful for judging the practical and technical merits of the tests under consideration.

Now judge the relative importance of these features and assign importance ratings, or weights, to them. We recommend a three-level weighting system like the following:

3=most important
2=average importance
1=useful, but not so important

The later, overall rating of a test is influenced by the importance weight of each feature. The purpose of having exclusionary features for screening tests at first and then importance weights for adjusting the influence of features on the overall rating is this: It is necessary to keep the less important features from adding up in the final analysis to overcompensate for the absence of essential and more important ones. In other words, don't let the minor test features dominate the comparison of tests. As noted above, a feature that is of minor importance for one test may be essential for a different use.

The different audiences and users of the tests in your program should participate in making the importance ratings so that their needs and interests will be taken into account. We recommend that teachers have a major voice at this stage because they have a good sense of how tests may or may not be useful for instructional purposes, of how practical a test is to administer, and of the effects of testing on pupils' motivation and morale.

STEP 3. Write the names of the tests to be compared at the top of the worksheet, and duplicate the form for the test rater(s).

In the spaces at the top of the worksheet, enter the name, form, and level of each test to be evaluated. For ease in filling out the rest of the worksheet, write an abbreviation of each test's name in the column labeled "Abbreviated Name."

Make a photocopy of the form for each person (or team of persons) who will be evaluating the tests, keeping the original copy blank in case more clean duplicates are needed.

MONTH/YEAR _____

RATER(S) _____

Step 3: NAMES/FORMS/LEVELS OF TESTS BEING COMPARED	ABBREVIATED NAMES	*Steps 8-10:* FINAL RESULTS TABLE (Total of Weighted Rankings for Each Test) 1st	1-2	2nd	2-3	3rd	3-4	4th	4-5	5th	Not Acceptable -Zero-

Step 1: TEST FEATURES	*Step 2:* IMPORTANCE WEIGHTS OF FEATURES 3=very imp. 2=important 1=useful	*Steps 5-7:* RANKINGS OF TESTS (Enter abbreviated names; for ties, average the respective ranks.) Acceptable: Best	Second	Third	Fourth	Fifth	Zero	NOTES

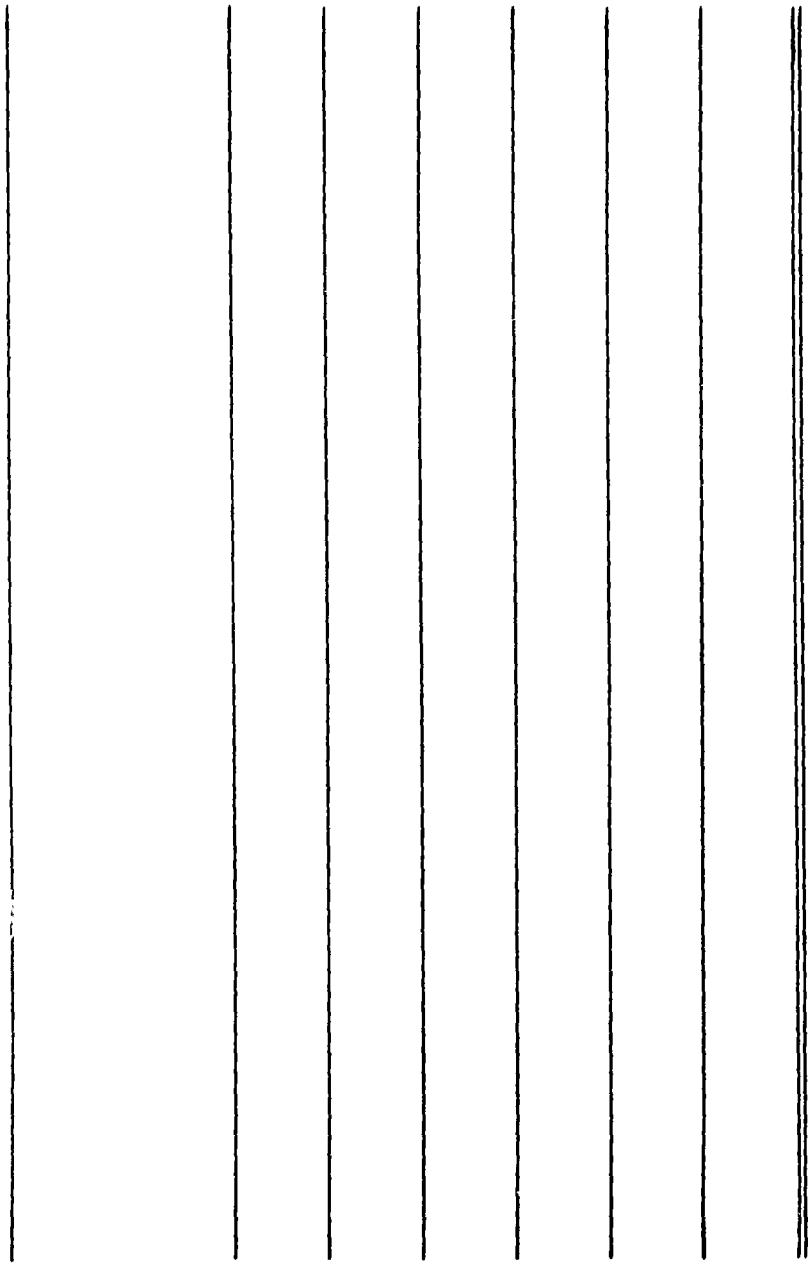


Figure 1. CSE Worksheet for Comparing Tests' Technical and Practical Features

MONTH/YEAR March 1979

RATER(S) Gee-Choy (except feature #8 - rated by evaluator)

Step 3: NAMES/FORMS/LEVELS OF TESTS BEING COMPARED	ABBREVIATED NAMES	*Steps 8-10* FINAL RESULTS TABLE (Total of Weighted Rankings for Each Test) 1st	1-2	2nd	2-3	3rd	3-4	4th	4-5	5th	Not Acceptable -Zero-
Test A (primary level)	A	~~HHT~~ ~~HHT~~ //	//	///							//
Test B	B	~~HHT~~	//	~~HHT~~ /							~~HHT~~ /
Test C	C			~~HHT~~ /							~~HHT~~ ~~HHT~~ ///

166

Step 1: TEST FEATURES	*Step 2:* IMPORTANCE WEIGHTS OF FEATURES 3=very imp. 2=important 1=useful	*Steps 5-7:* RANKINGS OF TESTS (Enter abbreviated names; for ties, average the respective ranks.) Acceptable: Best	Second	Third	Fourth	Fifth	Zero	NOTES
*1. Domain descriptions	3	A	B				C	Its detailed objectives make Test A easier to teach toward.
2. Agreement (of items and their domains)	3	B	A				C	Judges sorted items for Test B; then wrote a domain description for each set. Stray items were dropped. Sets where judges' and original domain descriptions matched were kept.

8. Consistency of scores (reliability--should be rated by a testing person)	3	A	B		C	Reliability of the decision is more useful than reliability of total scores. Test C is too low.
10. Item review	3	A	C		B	
14. Curriculum cross-referencing	2	B			A,C	
16. Alternate forms	2	A/B 1.5			C	(1+2) ÷ 2 = 1.5
20. Decision rules	3	A	C		B	If the scoring of Test C were more objective, then its decision rules would be too

*The numbers here correspond with those in Table 1, pages 14-15.

Figure 2. Worked Example of CSE Worksheet for Comparing Tests' Technical and Practical Features

179

STEP 4. Find, in the sample materials for each test, the evidence for the first test feature.

The specimen sets for many tests have an examiner's manual, a technical report, one complete test form for each test level, a complete set of answer sheets (if they are separate from the test forms), a complete set of scoring keys, examples of score reports, and any relevant stimulus cards, manipulanda, etc. Not all specimen sets are organized the same, and the evidence for any given test feature may be spread over several places.

The test rater should become familiar with the specimen sets, finding and noting the evidence for each feature which (s)he has the job of evaluating. If there appears to be no evidence for a given feature, that fact will be noted in the next step.

Find the evidence for the first test feature in all of the specimen sets.

STEP 5. Arrange the tests in descending order of merit on the given feature. Record these rankings (best, second, third...) in the respective columns of the worksheet next to the name of the feature.

Study the various tests' evidence for the given feature and decide which one (if any) is better than the others on that one dimension. Then decide whith test is second best, and so on. For any tests which provide no evidence of merit on a feature, or else evidence of insufficient merit, rank them as *zeroes* on that characteristic. You will have to decide locally how little merit a test can have on a feature and still be worth a ranking above zero. For example, you may decide that reliabilities below .60 are as bad as having no reliability data at all. Then you would rank all tests with no reliability figures or with figures below .60 as zeroes, and give the remaining tests positive rankings.

For this first feature, write the tests' abbreviated names in the columns for their respective rankings. Make these entries on the same line as the name of the feature. Be sure to write the short names of the zero-rated tests in the zero column because this information is used later.

STEP 6. For tests which are equally good on a feature, give them the average of the ranks they would have earned if not equal. For tests which differ, but not by much, use the given rules of thumb.

Occasionally two or more tests will be equally good on a given feature so that they are tied in ranking. For these cases, it is necessary to have a standard method of recording the rankings. A method that is commonly used with such ordinal (rank order) data is to assign each of the tied tests the *average* of the ranks they would have occupied if they had not been tied.

An illustration of this appears on the worked example in Figure 2. For feature #16 (alternate forms), Tests A and B have received equal ratings of 1.5 (1+2=3÷2=1.5). On the line of the worksheet for that feature, a circle has been drawn that includes the spaces for the first and second places. The abbreviated names of the two tied tests have been written in the circle as has the rating of 1.5.

In the same vein, if three tests were tied for third place, you would circle the spaces for third, fourth, and fifth, write the tests' short names in the circle, and write in the average of 3, 4, and 5, which is 4.

In short, give each of the tied tests the *average* of the ranks which they would have earned if not tied.

A related difficulty in ranking tests arises when they differ, but only slightly, in their merits on a given feature. Here you need to decide, "How much of a difference in quality *makes* a difference?" One rule of thumb is that small differences in merit should result in different rankings for test features that are very important, but not for features that are less important. A second rule of thumb is that small differences in merit should result in the same ranking for features that are judged subjectively or on which different judges disagree a great deal. For features on which clear, objective determinations can be made, there is justification for assigning different rankings on small differences.

You will still have to decide locally how much of a difference in quality should be treated as an effective difference, but the two rules of thumb will make those decisions easier.

STEP 7. Repeat Steps 4-6 for all other test features to be evaluated.

Compare the tests one feature at a time, and record their rankings for a feature before going on to evaluate the next one. When problems or questions arise, note them in the right column of the worksheet. They can be resolved later by conferring with other test raters or consultants. The "Notes" column can also be used to record reasons for a given ranking.

Staff members with special expertise should be assigned specific features to evaluate, so one person will not be rating all of the features. For example, language specialists will evaluate the linguistic and cultural appropriateness of a test for a bilingual program; testing specialists will rate the statistical features, etc.

STEP 8. Summarize the tests' rankings by weighting them and then recording them in the "Final Results Table" at the upper right of the worksheet.

Start with the rankings of the first feature. For the test that is ranked *Best*, you will enter one, two, or three tallies in the first column of the "Final Results Table" according to whether the feature has an importance rating of 1, 2, or 3. That is, the test which is ranked *Best* on a *Very Important* feature will have three tallies entered in the 1st place column of the table. Two tests that are tied for second and third place on that feature (hence are both ranked 2.5) will each have three tallies entered in the column headed 2-3 of the "Final Results Table." Any other fractional rankings will be transferred to the in-between columns of the summary table. Another test which had no acceptable evidence for that same feature would have three tallies entered in the right hand column of the table. All tallies will be written on the line of the table opposite the respective tests' name.

STEP 9. Check that the total number of tallies per test in the "Final Results Table" is equal.

Check your entries in the "Final Results Table" by counting the number of tallies for each test. The total number of tallies should be the same for each test, and should equal the sum of the importance weights for the features which were evaluated. If this is not so, re-do Step 8 on a sheet of scratch paper column by column, instead of feature by feature. Again verify your work by seeing if the number of tallies is equal and correct.

The outcome of this step is a table of profiles for the tests showing how many first places, in-between first and second places, second places, etc., each test earned. It is these overall profiles which will be compared next as the index of tests' technical and practical quality.

STEP 10. Compare the tests' profiles in the "Final Results Table." Eliminate tests that are markedly worse. Select the better ones for detailed analysis of their congruence with the local curriculum.

Refer now to the "Final Results Table" to decide whether any of the tests under consideration are markedly better or worse in their overall rankings. Either the profile of tallies for each test may be compared, or the tallies may be converted to percentages if percentages are easier to understand.[5]

Now compare the tests. Better tests have a greater part of their weighted ranks in the higher places, toward the left of the "Final Results Table." Tests of relatively lower quality and merit have a greater balance of their rankings in the Zero and other lower scores. Small differences between tests in the balance of high and low ranks should not be seen as significant, since the data do not come from precise measurement. At this stage of test selection, the purpose is to screen out tests that have markedly lower quality on the features which are relevant for your program.

If there is not an obvious break between the higher ranking and lower ranking tests, you may select and screen on the basis of your resources for carrying out the next step in test selection. That step involves studying tests item by item and judging the items' relevance to your curriculum. Since this analysis is quite detailed, you will want to carry it out on only a small set of tests. That consideration might lead you to select, say, the three top ranking tests in the "Final Results Table" for detailed curricular analysis. Retain the other tests in case the top three turn out to have too little relevance to your program.

[5]To transform the tallies into percentages, simply divide the total number of possible tallies (found in Step 9) into the number of tallies in each cell or box of the table. Record the numbers. The resulting figures are percentages of the total possible tallies which fall in each box. Adding across for each test, the percentages should sum to 100% (plus or minus rounding error).

SUMMARY

The methods in this chapter are meant to help you find, screen, and evaluate tests to suit your special situation. The complex judgment about the relative quality of tests is approached systematically by breaking it into a number of simpler judgments, then combining the results. Since these procedures are judgmental and not precise, you should regard them as hints for comparing tests, not as hard and fast rules. Feel free to adapt them to your needs and resources.

The most important aspect of tests, their relevance to the local curriculum, remains to be evaluated at this point. Chapter 6 takes up this final step in selecting a test.

CHAPTER 6
How To Select Tests: Comparing Tests For Their Relevance to a Given Curriculum

The previous chapter contained instructions for screening tests according to their potential uses and their technical merits. The measures which remain after screening can now be evaluated for their responsiveness to the local curriculum. In this chapter, procedures are described for rating the importance, content relevance, and difficulty of the objectives covered by a test, then comparing the ratings of the various tests. Three indices for comparing a test's congruence with the program are described: an overall measure, the proportion of a program's objectives covered by the test, and the proportion of the test's items that are relevant to the program.

INTRODUCTION

Achievement tests should be chosen so as to be maximally relevant to the test user's program. If the match between test and program is poor, then the test scores will not be useful for diagnostic or prescriptive purposes. Nor will such scores be useful for accountability or program evaluation purposes. Tests with low relevance to a given curriculum will not give fair credit for the successful teaching and learning which occur.

The research reviewed in Chapter 2 strongly supports the conclusion that tests have "curricular biases" affecting students' scores according to how well or badly the objectives tested match the objectives taught. Care taken in selecting tests for their curricular relevance will be rewarded when the scores are useful for instructional decisions and when evaluation results give credit for the program's actual achievements.

This chapter gives step-by-step procedures for comparing tests' curricular relevance. The procedures involve making a series of judgments about program objectives and test materials, expressing these judgments as numbers, combining the numbers for a single test, then comparing the results across tests. Table 3 gives a checklist of the steps for evaluating curricular relevance.

Because the method described in this chapter is a detailed one, you may wish to employ it only for major test selection decisions. Questions that may help determine whether a test selection decision is a major one include these:

- How many students will be tested?
- How much class time will be required for testing?
- Will the selected test be used repeatedly?
- Will the test's results be highly visible (e.g., to the public and to higher authorities)?
- Will the test results be used for decision making (e.g., about students, curriculum, teachers, or budget)?

The complexity of testing, both in terms of its relation to curriculum and in terms of numbers of people affected, requires the test selector to be very thorough and careful. In choosing a multilevel testing system, it is advisable to have each separate level of the test rated by teachers and curriculum specialists who are familiar with your program as it is actually taught. The objectives of most test batteries vary somewhat from level to level in content and in difficulty, so their appropriateness for your program may vary across levels as well.

The methods in this section ask you to compare test items with program objectives. There are several reasons for carrying out such a thorough analysis of tests before choosing one. First, these procedures help you to find the test that is most responsive to your purposes. Many tests are likely not to match your program well. Second, the procedures are explicit and easy to adapt to the constraints of your situation if you find yourself without sufficient time or resources to follow them exactly. Third, these procedures call attention to some aspects of tests which should not be overlooked, for example, the proportion of a test battery that is locally relevant, the proportion of the local curriculum which a test battery covers, the importance of the objectives covered, and the appropriateness of the test's difficulty for the program's students. Finally, the process of making numerical ratings at each stage of judgment and carrying them to the next stage ensures that information from earlier judgments is not forgotten or lost. As in the methods of Chapter 5, the component decisions all have an influence on the final rating of a test.

The methods described below deal with instructional objectives and with test items. Not every staff member is equally suited to use these methods. A number of educators are opposed to instructional objectives for various reasons. Many others do not have the patience or the style of thinking to deal with objectives. The best people for this task would not only be very familiar with the curriculum at the relevant level, but also have some skill in writing and recognizing objectives and a belief in the importance of curricular relevance in tests.

NOTE: In the following discussion, the word *skill* will sometimes be used interchangeably with the word *objective*.

TABLE 3
Checklist of Steps for Comparing
Tests' Relevance to a Given Curriculum

	Step		Page
____	1	Prepare a listing of the objectives of the program component to be tested.	176
____	2	Write your listing of program objectives to be tested in Column 1 of the Test Relevance Rating Form, called the worksheet (Figure 3).	178
____	3	Record the number of program objectives in Box B on the final page of the worksheet.	178
____	4	Rate the importance of each program objective in your listing, and record these judgments in Column 3.	186
____	5	Duplicate the worksheet for all of the raters and all of the tests still under consideration. Fill in the identifying information for each test to be rated.	186
____	6	List/index all of the items on the test in Column 2, each on the same line as the program objective that is most closely related to it.	186
____	7	Record the number of items on the test in both Box A and Box C on the final page of the worksheet.	186
____	8	Judge how closely the test items correspond with the respective program objectives in format, content, and process; record these judgments in Column 4.	187
____	9	Rate the appropriateness of the difficulty of each test item, and record the ratings in Column 5.	190
____	10	For each program objective that has any acceptable test items, multiply the ratings in Columns 3, 4, and 5 for each item; record the products in Column 6.	191
____	11	Add all of the products from Step 10, and record the sum at the bottom of Column 6 and in Box A.	191
____	12	Record the number of acceptable test items in Box C.	192
____	13	Compute the summary indices of tests' congruence with the curriculum, and record them at the bottom of the last page and the top of the first page of the worksheet.	192
____	14	Compare the summary indices of the tests under consideration. Decide whether one test has markedly greater congruence with your curriculum.	192

STEP 1. Prepare a listing of the objectives of the program component to be tested.

To find the test which is most relevant and responsive to your program, it is necessary first to be very clear about the instructional objectives of the curriculum to be tested. Such clarity is attained by making an explicit listing or index of these objectives. The listing should be prepared carefully, for it will serve as the standard of curricular relevance with which test materials will be compared.

Preparing such a list may be complicated if there are differences between the operational classroom curriculum and the official, formal one. Another complication arises when the operational curriculum varies from one organizational unit to another (i.e., from class to class or site to site). If there is little commonality of objectives from unit to unit, it will not be possible to draw up a realistic single listing. In this case, a single test cannot give a responsive, representative measure for all units, and the quick screening method of determining curricular relevance (Chapter 5) may be the best you can do.

Suggestions are given here for drawing up your list of curricular objectives under two conditions:

- When each subject area to be tested in the program has a uniform curriculum (even if there is a discrepancy between the operational curriculum and the official, formal one);
- When the objectives for the given subject area vary from organizational unit to unit, but there is great commonality in the important objectives.

1A. When there is a uniform curriculum, list (or index) the objectives for the program component to be tested as follows:

(1) Write the objectives in enough detail so that later in the process it will be possible to judge with confidence how closely a given test item measures or matches an objective. If, for example, your program teaches division in working (i.e., radical) form, but a test gives its division problems in number sentence form, your listing of local math objectives should enable the test rater to detect this difference and judge its importance. In the same vein, the listing of your language arts curriculum should enable the test rater to judge how well the words on a vocabulary test correspond with the vocabulary words in your program. Since curricular objectives are often stated rather generally, it will often be necessary to refine these objectives in order to use them as a basis for judging relevance of test items.

(2) When it would be burdensome to prepare such a full statement of your curricular objectives, an alternative is to prepare an index of them in the form of page references to the relevant teaching and exercise materials used in the classroom. For each separately teachable and testable skill, list in one place all of the pages where the skill is taught and practiced. A name or other verbal label for each of these skills should accompany the page references. This page referencing of skills to teachable materials will enable test raters to compare test items directly with instructional content and activities --a later step in the curriculum-matching process.

The referencing method of listing local curricular objectives may be

used either with or instead of the detailed method in 1A(1) above.

(3) In either instance above, it will help test raters to work with the listing if related objectives are grouped together. For example, a listing of fifth grade math objectives could be grouped under such headings as geometry, measurement, money, time, graphing, word problems, basic operations, and the like. For elementary reading, objectives could be grouped under such headings as phonics, structural analysis, sight words, vocabulary, comprehension, and the like. Subheadings can be used for smaller clusters of skills such as for the different basic arithmetic operations or the different types of comprehension skills which the curriculum covers. See Figure 4 for examples of subheads for grouping objectives.

(4) When the local curriculum is very detailed, your task of preparing a list of objectives can be simplified by combining small objectives. For example, if there are separate objectives for aural decoding of each speech sound in each of three positions within words--initial, medial, and final--this set of over 50 objectives could easily be reduced to six objectives dealing with consonants and vowels in each of the three positions. These six broader objectives would then be written in the listing instead of the many smaller ones. By combining very small, but closely related objectives, you can simplify the task of matching tests with curriculum without overlooking the more general skills which the specific skills comprise.

Two cautions should be noted regarding combining objectives. First, the amount of combining that is useful will vary with the intended use of the test. Combining will be of greater use for selecting survey tests than for selecting a battery of continuous progress tests. In the latter case, very detailed objectives, corresponding to individual lessons, might be needed. Second, it is possible to group too much. When objectives are broad and vague (e.g., *critical thinking*, *word attack*), their descriptions or labels do not make it clear what is being taught, learned, or tested. Such broad spectrum objectives do not describe the program skill in enough detail to allow the test rater to judge whether the relevant items measure the skill *as it is taught*.

(5) In cases where the formal, official curriculum and the operational classroom curriculum differ to any great degree, you will have to decide how to treat the differences. If the formal curriculum has not kept up with advances in classroom teaching, then it is reasonable to use the page referencing method in listing the program objectives. If, however, the formal curriculum accurately represents current program intentions, it is reasonable to follow the official formal objectives in preparing the listing. Other differences will need to be resolved on an individual basis.

1B. <u>When the operational, classroom curriculum varies from site to site, but there is great commonality in the important objectives for the program component to be tested, make a listing of the common objectives as follows</u>:

(1) Either compare listings of the separate classroom curricula and make a program listing out of the objectives that are common to the separate lists; or

(2) Give teachers of the different classroom level curricula a comprehensive listing of possible objectives for the appropriate level and subject. Ask the teachers to examine the master list and to check off the objectives which they actually teach at that level. Make a single program-wide listing out of the most commonly checked skills.

(3) Then go through the steps in 1A above to make this listing explicit, usable, and manageably short.

STEPS 2 and 3. Write your listing of program objectives to be tested in Column 1 of the Test Relevance Rating Form; record the number of objectives in Box B.

Contained in this chapter is a worksheet on which you can record the appropriate information as you follow the rating procedures. A blank version (Figure 3) and a worked example (Figure 4) of the worksheet are provided on the following pages.

Column 1 of the worksheet will contain your listing (or indexing) of the curricular component to be tested. This listing will be organized so that related objectives are grouped together under a common heading. Some of the smaller, more detailed objectives in your program may not appear separately in the listing because they have been grouped together into larger objectives.

Several sheets may be needed for listing or indexing the program component to be tested. *Number the pages and draw a heavy line under the last program objective, writing END OF LISTING in bold letters. Count the number of objectives in Column 1, and enter this number as the denominator in Box B on the final page of the worksheet.* Count only the objectives and not the names of curricular subareas or skill clusters. In Figure 4, there are 10 program objectives listed.

First sheet of _____ sheets

Step 5 | TEST NAME, LEVEL, AND FORM _____ RATER _____

PROGRAM SUBJECT AND LEVEL _____ DATE _____

OVERALL RATINGS: (fill in last)

GRAND AVERAGE_____ (average congruence per item ranging from 0-6)

INDEX OF COVERAGE_____ (proportion of program objectives measured by test)

INDEX OF RELEVANCE_____ (proportion of test that is relevant to program objectives)

Step 2 Listing of Program Objectives	*Step 6* Index of corresponding test items	*Step 4* Importance of program objectives	*Step 8* Match between items and objectives	*Step 9* Appropriateness of item difficulty	*Step 10* Combined judgments	Notes
		1=minor 2=important 3=essential	*0=not acceptable 1=adequate 2=very close*	*0=too hard or too easy 1=acceptable*	*Products across columns 3, 4, 5*	

Continuation sheet: page _____

TEST NAME, LEVEL, FORM _____ RATER _____ DATE _____

Listing of Program Objectives	Index of corresponding test items	Importance of program objectives	Match between items and objectives	Appropriate-ness of item difficulty	Combined judgments	Notes

Final sheet: page _____

TEST NAME, LEVEL, FORM _____ RATER _____ DATE _____

Listing of Program Objectives	Index of corresponding test items	Importance of program objectives	Match between items and objectives	Appropriateness of item difficulty	Combined judgments	Notes
					Step 11	
	Clearly irrelevant items				Sum of numbers in sixth column--enter in Box A also	

OVERALL RATINGS *Step 13*

BOX A	BOX B	BOX C
GRAND AVERAGE _____	INDEX OF COVERAGE _____	INDEX OF RELEVANCE _____
Sum of numbers in Column 6 (*Step 11*) _____ *divided by* Total number of test items (*Step 7*) _____	Number of program objectives adequately measured by test _____ *divided by* Total number of program objectives in Column 1 (*Step 3*) _____	Number of acceptable test items (*Step 12*) _____ *divided by* Total number of test items (*Step 7*) _____

Figure 3. CSE Test Relevance Rating Form

First sheet of 4 sheets

Step 5 | TEST NAME, LEVEL, FORM All American Test of Reading Comprehension, brown level | RATER Marion Choy

PROGRAM SUBJECT AND LEVEL 5th/6th grade reading comprehension | DATE 1/15/xx

OVERALL RATINGS*: (fill in last)
GRAND AVERAGE 2.1 (average congruence per item ranging from 0-6)
INDEX OF COVERAGE .70 (proportion of program objectives measured by test)
INDEX OF RELEVANCE .63 (proportion of test that is relevant to program objectives)

Step 2 Listing of Program Objectives	*Step 6* Index of corresponding test items	*Step 4* Importance of program objectives	*Step 8* Match between items and objectives	*Step 9* Appropriateness of item difficulty	*Step 10* Combined judgments	Notes
		1=minor 2=important 3=essential	*0=not acceptable 1=adequate 2=very close*	*0=too hard or too easy 1=acceptable*	*Products across columns 3, 4, 5*	
WORD LEVEL OBJECTIVES *(curricular subarea)*						
Word attack *(skill cluster)*						
1-Affixes: In a list of words --some of which have prefixes, some others of which have suffixes, and some of which do not have affixes--pupils will underline the affixes. The affixes will be drawn from this list: re-, pre-, un-, mis-, dis-, -ness, -less, -ful, -ly, -y, -en, and -er (as in driver).	p. 1, #1	2	2	1	4	
	2	↓	2	1	4	
	3		2	1	4	
	4		2	1	4	
	5		2	1	4	
	6		2	1	4	
2-Compound words: Pupils will complete compound words by matching words in a left column with words in a right column.	p. 2, #7	1	0 - - -	- - - - - - - -	- - - - - - - -	FORMAT WAY OFF--NOT SIMILAR ENOUGH TO PROGRAM OBJECTIVE
	8	↓	0 - - -	- - - - - - - -	- - - - - - - -	
	9		0 - - -	- - - - - - - -	- - - - - - - -	
	(continue next sheet)					

*Note: These ratings will vary with your judgments of your pupils' abilities and the importance of the program objectives.

TEST NAME, LEVEL, FORM All American Test of Reading Comprehension, brown level RATER Marion Choy DATE 1/15/xx

Listing of Program Objectives	Index of corresponding test items	Importance of program objectives	Match between items and objectives	Appropriateness of item difficulty	Combined judgments	Notes
	p. 2, #10	1	2	1	2	
	11	↓	2	1	2	
	12		2	1	2	
	13		2	1	2	
	14		2	1	2	
3-Root words: Given a list of words, each containing an affix, the pupil will write the root word. Affixes will include verb markers for tense and progressive, comparatives, and superlatives, and the ones for the objective on affixes above.	p. 3, #15	1	0 - - -	- - - - - - - -	- - - - - - - -	MORE IMPORTANT AS WRITING SKILL THAN READING ONE. ITEM FORMAT CALLS FOR PUPILS TO SELECT ROOT WORDS FROM FOUR VERY DIFFERENT CHOICES.
	16	↓	0 - - -	- - - - - - - -	- - - - - - - -	
	17		0 - - -	- - - - - - - -	- - - - - - - -	
Meaning						
4-Synonyms: Given a vocabulary word, the pupil will select from multiple choices the word or phrase which is a synonym.	p. 4, #18	2	2	1	4	
	19	↓	2	1	4	
	20		2	1	4	
5-Autonyms: Given a vocabulary word which has an opposite, the pupil will select its antonym from multiple choices.	p. 4, #21	2	2	1	4	
	22	↓	2	1	4	
	23		2	1	4	
Phrase, sentence, and text level objectives						
6-Meaning from context--words with one familiar meaning: Given sentences with one	p. 5, #24	1	2	1	2	
	25	↓	2	1	2	
	26		2	1	2	

TEST NAME, LEVEL, FORM All American Test of Reading Comprehension, brown level

RATER Marion Choy DATE 1/15/xx

Listing of Program Objectives	Index of corresponding test items	Importance of program objectives	Match between items and objectives	Appropriateness of item difficulty	Combined judgments	Notes
word omitted, pupils will select from multiple choices the one word whose meaning is most closely related to the context. Choices will be about the same length (± 2 letters) and at least two of them will start with the same letter.						
7-Meaning from context--words with more than one familiar meaning: Given sentences with a multiple-meaning word underlined, the pupil will pick from multiple choices the definition of the word which fits the context.	p. 5, #27 28 29	3 ↓	1 1 1	0 - - - 0 - - - 0 - - -	- -	TOO HARD
8-Main idea: Given a story of 3-5 sentences, pupils will select the main idea, where the three distractors deal with particulars of the story or with generalizations from single particulars.	p. 6, #30 31 32	3 ↓	2 2 2	1 1 1	6 6 6	
9-Inferences: Given a story in about three paragraphs, pupils will mark whether each of several supposed inferences from the story	p. 7, #33 34 35	1 ↓	2 2 2	1 1 1	2 2 2	THE SMALL DIFFERENCE BETWEEN ITEM AND PROGRAM PARAGRAPH LENGTH DOES NOT SEEM IMPORTANT.

TEST NAME, LEVEL FORM All American Test of Reading Comprehension brown level

Final sheet: page 4

RATER Marion Choy DATE 1/15/xx

Listing of Program Objectives	Index of corresponding test items	Importance of program objectives	Match between items and objectives	Appropriateness of item difficulty	Combined judgments	Notes
is probably true, probably false, or can't tell.						
10-Meanings of colloquial phrases: Given a sentence with an idiomatic colloquial phrase underlined, pupils will select the literal phrase with the same meaning from multiple choices.	p. 8, #39 40 41	1 ↓	2 2 2	0 - - - 0 - - - 0 - - -	- -	TOO EASY. THE DISTRACTORS DON'T MAKE SENSE SO THEY COULDN'T BE CORRECT CHOICES.
END OF LISTING	p. 7, #36, 37, 38 Clearly irrelevant items				*Step 11* 88 Sum of numbers in sixth column--enter in Box A also	

OVERALL RATINGS *Step 13*

BOX A	BOX B	BOX C
GRAND AVERAGE 2.1	INDEX OF COVERAGE .70	INDEX OF RELEVANCE .63
Sum of numbers in Column 6 (*Step 11*) 88 *divided by* Total number of test items (*Step 7*) 41	Number of program objectives adequately measured by test 7 *divided by* Total number of program objectives in Column 1 (*Step 3*) 10	Number of acceptable test items (*Step 12*) 26 *divided by* Total number of test items (*Step 7*) 41

Figure 4. Worked Example of CSE Test Relevance Rating Form

STEPS 4 and 5. Rate the importance of each program objective. Duplicate the worksheet and fill in the identifying information for each test to be rated.

In Step 4, judgments are made about the importance of each of the objectives that is listed in Column 1. These judgments are then expressed in numbers, indicating degrees of importance, and are recorded in the third column.

For each of the program objectives, the test rater is to judge how important it is for students to attain. The number of degrees of importance you decide to use is a matter of local judgment, but three degrees (minor, important, and essential) offer a balance of convenience and contrast.

For each objective that is judged to be of minor importance, assign it a rating of 1, and record the rating in the third column on the same line as the objective. A minor objective is one that could be omitted with little harm to student progress. Important objectives, ones that clearly contribute to progress or are worth learning for their own sake, are assigned a rating of 2. Essential objectives, ones that are prerequisites or are essential for student progress, are given a value of 3.

After judging the importance of each program objective and recording its importance rating in Column 3, check the ratings by comparing them with one another. That is, after judging all objectives separately, confirm the ratings by seeing if ratings seem appropriate relative to one another.

On completing all of the steps up to this point, make enough copies of the partially filled-in CSE Test Relevance Rating Form to permit all of the raters to rate all of the tests under consideration. Keep the original form blank in case more copies are needed. For each test, fill in the blanks at the top of each page of the worksheet.

STEPS 6 and 7. List/index all of the items on the test in Column 2, each on the same line as the program objective that is most closely related to it. Record the number of test items in both Box A and Box C on the final page of the worksheet.

Look at each test item and decide which program objective in Column 1, if any, it seems to measure. For each item, write its number (or test page and number) in Column 2 opposite the relevant program objective. At this stage, be generous in judging whether an item is responsive to an objective; what is important here is to assemble for each objective all of the items that measure it, even remotely.

Try to pair each test item with only one program objective; but if an item seems to measure more than one program objective, write its number in Column 2 opposite each objective. Circle any repeated listing of a single item for later reference.

There will probably be some items on the test which do not correspond to any of the objectives in Column 1. List these items at the end of Column 2, next to the END OF LISTING in Column 1. Enter either the item number or page and number so that you and other test raters can compare your judgments about the items.

Ideally, you would be able to list or index a test's objectives rather than its items in Column 2 next to the relevant program objectives. In

fact the objectives of existing tests are not specific enough to serve as a basis for judging test relevance accurately.

Before going on, count the total number of items on the test being rated, and enter the number as the denominator in Boxes A and C on the final pages of the worksheet. If you make this tally by counting numbers in Column 2, make sure not to count any item more than once. That is, do not count any circled (i.e., repeated) items.

STEP 8. Judge how closely the test items correspond with the respective program objectives in format, content, and process; record these judgments in Column 4.

The purpose of this step is to judge how relevant or sensitive each item is to the corresponding objective that your program teaches. Examine each test item, and judge how closely it corresponds to the respective program objective in *format*, *content*, and *process tested*. The correspondence may be not acceptable, adequate, or very close. For these degress of match/mismatch, assign a score of 0, 1, or 2 respectively and record it in the fourth column.

If the item format (e.g., matching pictures and words) differs from the format of the relevant instruction and practice, decide whether that difference will interfere with your students' displaying their learning of the program skills on the test. If the test format is so unfamiliar as to make it very hard for students to show their learning of the program skill, then a zero rating should be recorded.

Attend also to the content and process which the item measures. For objectives dealing with specific knowledge (e.g., vocabulary), make your judgment according to how closely the content of the item samples the content of the instruction. For objectives dealing with processes (e.g., identifying the main idea), decide how well the process, as taught, matches the process needed to answer the item correctly.

Record the overall rating of format, content, and process in Column 4 as one number. For an item earning a zero rating, draw a horizontal line through the next two columns to show that it does not need to be rated further.

The issue of program and item *content* is illustrated by comparing the first program objective in the worked example with sample test items 1-6 in Figure 5. The objective reads as follows:

> *Affixes: In a list of words--some of which have prefixes, some others of which have suffixes, and some of which do not have affixes--pupils will underline the affixes.*

Sample test items 1-6 (in Figure 5) earn a congruency rating of 2 in the worked example in part because their content is completely congruent with the program objective on affixes.

For such judgments, you may need to set some arbitrary criteria such as the following:

- 90-100% congruence between item content and content described by the program objective rates a 2
- 80-90% congruence rates a 1
- < 80% congruence rates a 0 as not acceptable

DIRECTIONS: In the list of words below, draw a line under each prefix or suffix. Some of the words do not have a prefix or a suffix. A worked example is given in the box.

EXAMPLE:

rewrite
happy
watchful

Draw a line under each prefix or suffix.

1. dislike
2. during
3. driver
4. people
5. quickly
6. refill

DIRECTIONS: Read each group of four words below. If all four words are compound words, circle Yes. If any word is not a compound, circle No. The first two are done for you.

EXAMPLE:

Inkblot, screwdriver, pigskin, notebook — (Yes) No

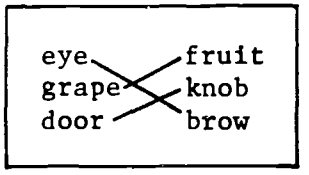
EXAMPLE:

Hammer, teamwork, keychain, enemy — Yes (No)

7. Afternoon, barefoot, walking, mailed — Yes No
8. Fireplace, football, bedtime, icebox — Yes No
9. Bookcase, ruler, raindrop, heavenly — Yes No

DIRECTIONS: In each box below, a word on the left makes a bigger word with one word on the right. Draw a line to connect the two words that make a bigger word. The first box is a worked example for you.

EXAMPLE:



Explanation
eyebrow
grapefruit
doorknob

10-14.

an	noon
after	fly
any	light
butter	body
flash	other

Figure 5. Sample test items

DIRECTIONS: Read the following sentences.

The next morning the two men came back for Brown Pet.
Jack and Nancy ran to the barnyard
They wanted to tell the cow good-by.
Mr. Stone said, "Your pet will be happy at the zoo."

If the sentence below could be *true*, check A. If the sentence is *probably false*, check B. If you *can't say* whether it is true or false, check C. The first question is done for you.

EXAMPLE:

The men were going to take Brown Pet away.

√a. Probably true
b. Probably false
c. Can't say

33. Brown Pet was in the barnyard.

a. Probably true
b. Probably false
c. Can't say

34. The men were taking Brown Pet to the zoo.

a. Probably true
b. Probably false
c. Can't say

35. The men came for Brown Pet in the morning because it would take all day to get to the zoo.

a. Probably true
b. Probably false
c. Can't say

Figure 6. Sample test items*

*Adapted from the Behavioral Objectives and Test Items bank, Glen Ellen, Illinois.

The issues of item *format* and item solution *processes* are illustrated by comparing the second program objective on compound words in the worked example with items 7-9 and 10-14 in Figure 5. The program objective reads as follows:

> *Compound words: Pupils will complete compound words by matching words in a left column with words in a right column.*

Items 10-14 fit that description. But items 7-9 present lines of four words and ask the student to circle *Yes* or *No* for each line. The latter format is different from the one used in the program and probably much less familiar.

Item format often affects the mental processes which a pupil must use for coming up with correct answers. In items 7-9, pupils need to be able to understand the concept of *all four words* and to keep it in mind while reading the words. They also need to break down each word in items 7-9, sometimes more than once, and judge whether each part is a real word:

- fi - replace
- fire - place

Some of the parts are real words and others are not. A student who uses an efficient method for doing these problems analyzes each word in the item until (s)he finds a non-compound. On finding a non-compound, s(he) will circle *No* and go to the next item directly. If all of the words in the item are compounds, the test taker circles *Yes* and goes on.

In contrast, the processes for solving items 10-14 involve remembering a word on the left, building possible compounds out of it with words on the right, judging each possible compound, continuing until a compound is recognized, and repeating the process until all of the words on the left are used.

If the difference between the program objective and the content/format/process of items 7-9 will interfere with your pupils' using their program skill to answer those items, assign a congruency rating of 1 or 0, depending on whether you judge the items to be acceptable reflections of the objectives, or unacceptable. In Figure 4, the differences in format and processes between sample items 7-9 and the program skill on compound words were judged to be unacceptable. Record the rating for each of the items in the column for Step 8 on the line where the respective items are indexed.

A second example of a difference between a program objective and a tested one occurs with the sample items on inferential comprehension in Figure 6. The program objective asks for stories which are about three paragraphs long. The items use a text which is rather short. If you think that the difference does not really change the objective, then you will want to assign a rating of 2 (very close) to the items and record it in the column for Step 8 on the lines where the respective items are indexed. If the difference in program and test text length does change the objective somewhat, then assign and record a lower congruency rating.

STEP 9. Rate the appropriateness of the difficulty of each test item, and record the ratings in Column 5.

The last judgment of test items involves rating the appropriateness of each item's level of difficulty. Difficulty judgments are expressed

on a two-point scale where 0=too hard or too easy, and 1=acceptable. These judgments are then recorded in the fifth column of the worksheet. It will help in making these judgments to ask yourself these questions:

- Is the item so easy that students who are unskilled on the program objective will answer it correctly much of the time?
- Is the item so difficult that students who have mastered the program objective will miss it much of the time?

Whenever the answer is *yes*, the item should get a zero rating. For all such items, draw a horizontal line through the next column to the right.

As in Step 8, these judgments require you to study the test items. It it proves hard to separate judgments of item difficulty from those of format and content (Step 8), then this fifth column can be eliminated and the overall task simplified by one step. Teachers and curriculum specialists who are very familiar with your program as it is actually taught will be able to make these two types of judgments simultaneously with confidence. Anyone who is not intimately acquainted with the operational curriculum will have trouble with the process.

An alternative to judging items' difficulty is to use the test publisher's field test data. This option is open only for tests which give item difficulty figures based on the responses of an appropriate comparison group of pupils.

STEPS 10 and 11. For each program objective that has any acceptable test items, multiply the ratings in Columns 3, 4, and 5 for each item; record the products in Column 6. Then add all of the products, and record the sum at the bottom of Column 6 and in Box A.

A total rating for each test item is now reckoned by multiplying the importance value of the respective objective (Column 3) by the item's ratings for curricular match (Column 4) and difficulty (Column 5). Items getting unacceptable ratings in Columns 4 or 5 will already have been lined out in Column 6.

The numbers in Column 6 are summaries of the test raters' judgments about the importance, curricular relevance, and difficulty of the objectives covered by a test. These numbers range in possible value from 1 to 6. A rating of 6 would be received by a test item that:

- Measures a very important program objective (rated 3 in Column 3)
- Matches the objective closely in content and format (rated 2 in Column 4)
- Has an acceptable level of difficulty (rated 1 in Column 5)

The overall rating for such an item then comes from multiplying across the form, 3x2x1=6 and is entered in Column 6.

After multiplying the ratings and recording them in the sixth column, check your arithmetic. Then add the numbers in this column, and record the sum at the bottom of the column. Also, write it in Box A as the numerator.

STEP 12. Record the number of acceptable test items in Box C.

As a step toward finding the proportion of the test's items which are relevant to your program, count the number of acceptable items. These items are the ones which were *not* lined out in Column 6 (Step 10). In other words, count the number of numbers in Column 6, and record it as the numerator in Box C on the last page of the worksheet.

STEPS 13 and 14. Compute the summary indices, and use them to compare tests' congruence with your curriculum.

To summarize a test's curricular relevance, three indices are computed: the *Grand Average*, *Index of Coverage*, and *Index of Relevance*. The Grand Average, which may range in value from 0 to 6, describes the average, per test item, of the combined judgments of importance (Step 2), curricular match (Step 8), and item difficulty (Step 9). Compute the Grand Average by dividing the result of Step 11 by the total number of items on the test (Step 7). Record this number in Box A on the final page of the worksheet.

The Grand Average for a single test takes on meaning when compared with the same figure for other tests. The one test with the highest Grand Average does a better job of covering more of the important program objectives. This one comparison still does not indicate whether the highest rated test covers the program well enough. That judgment is aided by two other statistics on the worksheet, the Index of Coverage and the Index of Relevance.

The Index of Coverage tells how completely a test covers the program objectives listed in the first column. It is derived by dividing the number of objectives in Column 1 (Step 3) into the number of those objectives which the test measures adequately. Adequacy of measurement is determined by two factors: the number of test items per objective and their goodness of match to the objective. Test raters will have to use their discretion in deciding whether the number of items measuring an objective is sufficient. This decision, however, will be guided by the intended use of the test. One or two good items per objective might be enough for a survey test, but eight to ten might be a minimum for a battery of tests for monitoring progress. In counting items per objective, count only the ones which have an acceptable match with the program objective, that is, which get a numerical rating in the sixth column of 1 or higher.

While the Grand Average is based on test items, the Index of Coverage is based on numbers of objectives: the proportion of program objectives (Column 1) that are *adequately* measured. Its possible values range from a low of zero to a high of 1.0. If the value of the Index of Coverage for one test is .6, then 40% of the program objectives to be tested are *not* covered by the test. For tests that differ very little on the Grand Average, the one with the highest Index of Coverage would be preferable. This summary statistic is recorded in Box B.

The last summary figure for comparing tests is the Index of Relevance, which tells what proportion of the test is sufficiently relevant to your program. It is computed by dividing the total number of items on the test (Step 7) into the number of items that adequately match the program (Step 12). Those items are the ones that receive a numerical

rating of 1 or higher in the sixth column of the rating form.

The Index of Relevance has possible values ranging from zero (totally unresponsive to the local program) to 1.0 (all of the test items are adequate measures of program objectives). On a test with a relevance rating of .75, a quarter of the items measure objectives that are either *not* part of your curriculum or are not at the right level of difficulty.

This third factor is important because selecting a test with a large percentage of items that are not relevant to your program means paying, both in time and money, for test materials that work against you. Your students may do poorly on objectives in the test which do not match your program, and the test results will not be very helpful for assigning program lessons. Enter the Index of Relevance figure in Box C.

Each of the three summary figures gives a different piece of information about a test. Since they are based on different types of information, it would not be meaningful to add them for a single summary judgment. The final choice of a single test will be based on a comparison, across several tests, of each of the summary figures. To facilitate this comparison, enter the three summary figures in the spaces provided at the top of the first page of the worksheet.

Other useful kinds of information can be derived from the CSE Test Relevance Rating Form. For example, the average importance of program objectives not covered in a test could be reckoned and compared as a supplement to the other three summary measures. Also, the entries in the sixth column of the worksheet can be used to guide the scoring and reporting of pupils' responses to a test. Items which are identified before the testing occurs as program-irrelevant can later be omitted from the analysis of scores. Total test scores could be reported, if required by higher authority, but the customized, program-relevant scores would provide an important context for interpreting the total scores.

ON INCREASING THE RELIABILITY OF THESE METHODS

The basis of the methods given in this chapter is human judgment, not precise physical measurement. These methods are an aid to judgment and memory, not an errorproof mechanism for measuring tests. Since the choice of tests is a social/political one which depends on knowledge of curriculum and pupils, it cannot be completely automated. These methods reduce the unreliability of judgment by providing some uniform rating scales (namely, importance of objectives, congruence of items with objectives, and difficulty of items) and uniform cutting points or criteria along these scales. Furthermore, the individual ratings are recorded as they are made and are combined in a uniform manner, rather than left unrecorded to be combined in an impressionistic and forgetful manner.

The users of these methods can increase their reliability further by several means. First, it will help to give the test raters some practice before having them do an operational comparison of tests' curricular relevance. A part of your program curriculum may be used, for familiarization, along with a

real test. Next, it will help for test raters to discuss with one another the judgmental scales for the purpose of encouraging uniformity in applying the cutting points to the scales.

Third, it is important to have each level of a test rated independently by more than one person. Where two or more raters disagree, they may resolve their differences, or they may decide that they have well founded differences of judgment and split the differences. A final, and essential, method for increasing the reliability of ratings is to have the job done for pay, not on your staff's time off. These methods are labor, not play, and they are a part of making your program function better.

Although the procedures in this chapter are detailed, they are easier to carry out than to read about. They are intended as a flexible prototype to be adapted to local needs and resources. The attention to detail will be rewarded by your choice of a test that comes closest to meeting your needs.

APPENDIX A
Resources for Developing CRTs Locally and for Purchasing Made-to-Order CRTs

Many school districts undertake to write their own test batteries to ensure that their unique testing needs will be met. There are several types of resources which can make local development of objectives-based tests feasible, if not easy.

The first are reference works on methods of item and test construction. These books are not specifically on criterion-referenced testing, but they are a great help in writing good test items. *Books in Print* lists such sources under subject headings like "educational tests and measurement." Second, there are works on creating CRT materials, a number of which are listed below.

Next, there are lists of objectives around which a curriculum, a continuum, or a testing battery may be built. Comprehensive sets of objectives in a variety of subject areas are sold separately from test materials by various publishers such as Commercial-Educational Distributing Services, Instructional Objectives Exchange, and Westinghouse Learning Corporation. In addition, many school districts have prepared curriculum guides or objectives lists which are uncopyrighted. A small sampling of these is included below.

The fourth resource for local test development is item banks, that is, pools of existing test questions. Along with the objectives lists, objectives-based item banks are listed below. The ones listed here are in the public domain and thus may be reproduced or modified locally. Even when the pre-existing materials are used only as models, they save much of the labor involved in thinking of possible objectives, selecting formats for test items, and developing distractors. The item banks listed in this Appendix are not included in the test reviews of this book because they did not meet all of the screening criteria.

Finally, there are publishers who provide made-to-order CRTs for purchase.

Sources on How To Develop CRT Materials

Baker, E. L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. *Educational Technology*, 1974, *14*(6), 10-16.

Gronlund, N. E. *Preparing criterion-referenced tests for the classroom.* New York: Macmillan, 1973.

Hambleton, R. K., & Eignor, D. R. *A practitioner's guide to criterion-referenced test development, validation, and test score usage* (2nd ed.), 1979. [Until these materials are published commercially, they are available from the Clearinghouse for Applied Performance Testing, Northwest Regional Educational Laboratory, 710 S.W. Second Avenue, Portland, Oregon 97204.]

Popham, W. J. *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice Hall, 1978.

Roberson, D. R. *Development and use of criterion-referenced tests.* Austin, TX: Educational Systems Associates, 1975.

Sherman, M., & Zieky, M. (Eds.). *Handbook for conducting task analysis and developing criterion-referenced tests of language skills.* Princeton, NJ: Educational Testing Service, 1974.

Sullivan, H. J., Baker, R. L., & Schutz, R. E. Developing instructional specifications. In R. L. Baker & R. E. Schutz (Eds.), *Instructional product development.* New York: Van Nostrand Reinhold, 1971.

Sweezey, R. W., & Pearlstein, R. B. *Guidebook for developing criterion-referenced tests.* ERIC Document TM 005 377, 1976.

Tombari, M., & Mangino, E. *How to write criterion-referenced tests for Spanish-English bilingual programs.* Austin, TX: Dissemination and Assessment Center for Bilingual Education, 1978. [Write DACBE, 7703 N. Lamar Blvd., Austin, Texas 78752.]

Objectives Lists and Banks of Objectives-Based Test Items

Behavioral Objectives and Criterion-Referenced Test Items in Mathematics, K-6.

Uniondale Public Schools, Uniondale Union Free School District, Uniondale, New York 11553.

For each of the grades, there are two pamphlets, one with 80 or more objectives, the other with an item bank for testing those objectives.

Cost: cost of copying complete set approximately $20.00.

Behavioral Objectives and Test Items:

Language Arts	(ERIC numbers ED 066 498 through 501)
Mathematics	(ERIC numbers ED 066 494 through 497)
Social	(ERIC numbers ED 066 502 through 504)
Science	(ERIC numbers ED 066 505 through 508)

Institute for Educational Research, 793 N. Main Street, Glen Ellen, Illinois 60137.

A bank of approximately 5,000 objectives and 27,000 accompanying test items was written by Chicago elementary and secondary school teachers in the course of their participation in workshops in the writing of behavioral objectives and test items. Objectives and items in each of the four content areas are available for primary, intermediate, junior high, and high school levels. A volume on measuring students' attitudes (ED 066 493) and an operational guide to the workshops (ED 066 492) are also available. Parts of the materials are also available through the Objectives and Items CO-OP, listed below. The Institute for Educational Research expects to have revised materials in the areas of math and language arts available in the fall of 1979 for purchase.

Behavioral Objectives Curriculum Guide, Mathematics, Grade 7.

Bucks County Public Schools, Routes #611 and #313, Doylestown, Pennsylvania 18901.

A framework for the development of a seventh grade mathematics program. The guide includes over 160 behavioral objectives with an assessment item and estimated learning time for each objective at three levels of difficulty.

Cost: $3.00.

Individualizing Mathematical Learning in the Elementary Schools:

An Ordered List of Mathematical Objectives, K-8
Test Items for Primary Mathematics
Test Items for Intermediate Mathematics

CCL Document Service, 1025 W. Johnson St., Madison, Wisconsin 53706.

Approximately 200 mathematics objectives are available along with 400 to 500 sample test items keyed to the objectives. The items and objectives were developed at the Wisconsin Research and Development Center at the University of Wisconsin in cooperation with the Wisconsin Department of Public Instruction. Although the test items have been out-of-print since mid-1976, single copies will be made upon request.

Cost: objectives, $1.00; primary items, $7.30; and intermediate items, $12.65.

Junior High Unified: Sequencing and Keying of Unified Studies; Test Specifications for Criterion-Referenced Testing; Achievement-Awareness Record for Language Arts. ERIC Document ED 116 193.

ERIC Document Reproduction Service, P.O. Box 190, Arlington, Virginia 22210.

This language arts curriculum guide for grades 7-9 was developed by the Shawnee Mission (Kansas) School District. It includes 50 objectives with sample test items on composition. Objectives without sample test items are given in the following areas (number of objective in parentheses): syntax (81), listening and viewing (20), literature and reading (24), and speaking (18).

Cost: $6.01 for hard copy plus 66¢ postage.

Managing Readin' by Objectives

El Dorado County Office of Education (attn: Curriculum Clerk), 337 Placerville Drive, Placerville, California 95667.

This is a reading skills management system developed by teachers and district staff in El Dorado County, California. The 1971 edition is a revision that was based on teachers' classroom experience with the system. The biggest component is a bank of over 10,000 items keyed to over 600 objectives in the following four skill categories: language development (oral and written language, vocabulary), word analysis (sight words, phonics, morphology), comprehension (ten types), and study skills (twelve subareas). Items are divided into eleven levels from pre-reading through grade 8. Many of the individual tests will have to be recopied before duplicating, for example, where fill-in items are already filled-in with the correct answer. Single copies of the item bank are sold as well as the optional resources for a complete testing and accountability system listed below. The manual includes informal measures for diagnosis.

Cost: The complete bank of Criterion Questions for all levels is $23.50. The manual/kit is $6.00; record sheets are 10¢ for individual pupils and 30¢ for the class chart; U-sort task cards are $36.00 per all-level set.

Mathematics Assessment Process Handbook of Objectives, K-9, 1973-74.

Greece Central School District, Greece, New York 14616. Available from the Clearinghouse for Applied Performance Testing, Northwest Regional Educational Laboratory, 710 S.W. Second Avenue, Portland, Oregon 97204.

Described as a minimum skill component of the total district mathematics curriculum, this system contains guidance on classroom management plus over 200 mathematics objectives, each with a sample item.

Cost: $13.40.

The Objectives and Items CO-OP:

Language Arts
Mathematics
Social Studies
Science
Vocational Education

The CO-OP, 413 Hills House North, University of Massachusetts, Amherst, Massachusetts 01002.

The CO-OP has collected over 10,000 objectives and 40,000 items for elementary and secondary school levels in 47 booklets developed by a number of school systems and state education departments for their own use. For example, the mathematics materials include those of Project SPPED developed for the New York State Education Department. Parts of the Behavioral Objectives and Test Items, listed above, are included in the CO-OP's materials. These booklets are described as varying in comprehensiveness and have not been edited by the CO-OP.

Cost: $1.00 to $58.00 per booklet; complete sets by content area--language arts, $141.50; mathematics, $384.50; social sciences, $58.00; science, $77.00; vocational education, $29.00

Phoenix Minimal Objectives:

Minimal Mathematics Objectives, K-12, 1975
Proposed Minimal Reading Objectives, K-12, 1974
Proposed Minimal Writing Objectives, K-12, 1974

Curriculum and Instructional Development Services, Greater Phoenix Curriculum Council, 2526 W. Osborn Rd., Phoenix, Arizona 85017.

This system contains over 300 basic skills objectives, each with one or more sample items or suggestions for the writing of assessment items or tasks. Broad performance tasks, involving several skills, are frequently suggested to evaluate mastery of writing objectives.

Cost: mathematics, $2.00; reading, $6.00; and writing, $2.50.

Sample Assessment Exercises Manual for Proficiency Assessment:

Volume I: Sample Exercises
Volume II: Item Statistics for Grades 7, 9, and 11
Technical Assistance Guide for Proficiency Assessment

Cashier, State Department of Education, 515 L Street, Sacramento, California 95814.

The first volume gives item specifications and a pool of about 1500 test questions for three models of proficiency assessment: school context (reading, writing, and math), functional transfer (forms, maps, ads, directions, and measures), and applied performance. Volume II gives item statistics for most of these items along with a description of the field test and directions for reading and using the statistics. The Technical Assistance Guide has a variety of resources for setting up a proficiency assessment program.

Cost: $54.00 for Volumes I and II together. No charge for Guide.

Names and Publishers of Made-to-Order CRTs

Comprehensive Achievement Monitoring (CAM)
National Evaluation Systems
P.O. Box 226
Amherst, Massachusetts 01002

Customized Criterion-Referenced Tests
Multi-Media Associates, Inc.
P.O. Box 13052
4901 E. Fifth Street
Tucson, Arizona 85732

Customized Objective Monitoring Service
Houghton Mifflin Company
777 California Avenue
Palo Alto, California 94304

IOX Test Development Service
Instructional Objectives Exchange (IOX)
P.O. Box 24095
Los Angeles, California 90025

Mastery Custom Tests - Reading and Math
Science Research Associates
259 East Erie Street
Chicago, Illinois 60611

ORBIT
CTB/McGraw-Hill
Del Monte Research Park
Monterey, California 93940

APPENDIX B
Sources of Other Test Reviews

Buros' *Mental Measurements Yearbook* (The Gryphon Press) is the most familiar source of test reviews. Recently a number of other books specializing in reviews of educational tests have been published. Among these, the following three were funded by the National Institute of Education:

- Hoepfner, R., et al. *CSE secondary school test evaluations.* Los Angeles: Center for the Study of Evaluation, 1974.

- Hoepfner, R., et al. *CSE elementary school test evaluations.* Los Angeles: Center for the Study of Evaluation, 1976 (2nd edition).

- Pletcher, P., Locks, N., Reynolds, D., & Sisson, B. *A guide to assessment instruments for limited English speaking students.* New York: Santillana, 1978.

The first two of these volumes deal exclusively with norm-referenced tests. Two other test review books were funded by the Office of Education, namely:

- *Tests of adult functional literacy.* Portland, OR: Northwest Regional Educational Laboratory, 1975.

- *Assessment instruments for bilingual education.* Los Angeles: National Dissemination and Assessment Center at California State University, Los Angeles, 1978.

A number of professional journals also carry reviews of tests that are relevant in educational settings:

- *Bilingual Resources*
- *Educational and Psychological Measurement*
- *Journal of Counseling Psychology*
- *Journal of Educational Measurement*
- *Journal of Special Education*
- *Review of Educational Research*

While the present volume was in preparation, articles comparing and evaluating specific criterion-referenced tests began to appear (Denham, 1977; Stallard, 1977a, 1977b; Hambleton and

Eignor, 1978). Articles of this nature are indexed under appropriate subject, author, and title headings in *Resources in Education* and *Current Index of Journals in Education*.

APPENDIX C
Glossary

This section gives definitions for most of the technical terms used in this book. The focus is on basic terms dealing with criterion-referenced testing, many of which are relevant also to norm-referenced testing. The definitions are designed to introduce basic concepts in a non-technical manner.

Tests exist for a multitude of objectives, traits, and behaviors. In this glossary, we use the summary phrase "test of a skill or attitude" to indicate a test of anything from maximum performance (such as knowledge, skill, or achievement) to typical performance (such as attitude or trait).

Absolute score	A test score reporting the number or percentage of items correctly answered (cf. *comparative information*).
Alternate form	A second version of a test with the same format, content, and difficulty as the first version, but with different test items. Tests with alternate forms may be useful for assessing learning with a pretest-posttest procedure. Pupils' scores on a second testing are more valid when a second form is used because those scores are less influenced by students' specific memory for the content of the form used for the first testing.
Amplified objective	A form of test specification which consists of a behavioral objective, sample test item, a description of possible item format, and a description of content that may be included in item stems and responses.
Assessment	The measurement of a thing's quality, amount, or effectiveness, such as an assessment of a student's learning.

Behavioral objective

A statement, usually in the following form,

Given (specific materials), *students will* (perform specified responses),

which describes an outcome of instruction in terms of a testing situation. It is called a *behavioral* objective because it describes not only the test content or subject matter, but also the observable behavior which the student is supposed to exhibit in responding. Examples of observable behavior--

select from multiple choice alternatives,
write a 300-word essay,
repeat aloud--

contrast with non-observable behaviors which are typical of more general educational goals--

know,
understand,
appreciate,
solve.

Bias

A flaw in test construction which causes the test scores to be unfairly influenced by the test takers' experience outside the classroom or by traits that are not responsive to experience in school.

Comparative information

Information that helps to interpret individual or group test scores by comparing them with the scores of other test takers. Some of the types of comparative information are percentiles, grade level equivalents, scores of criterion groups.

Conceptual validity

A term coined for this book which refers to aspects of a criterion-referenced test's validity that are not determined by field testing. These include the quality of the test specification, the match between the items and their specifications, and representativeness of the items.

Concurrent validity	The validity of a test whose scores correlate highly with contemporaneous criterion behaviors (cf. *criterion*[c]). For example, a pencil and paper test of skills in auto repair has concurrent validity if pupils who earn higher scores on it also are more proficient in the criterion skill of repairing autos.
Confidence interval	A statistical estimate of the interval within which a score, if it were error-free, would probably fall. Interval estimates contrast with single point estimates, such as the average, and are assigned probabilities which are called "levels of confidence."
Consistency	A general term used in this book for the various types of reliability. The term is used rather than the term *reliability* to call attention to the fact that measurement specialists disagree on the usefulness of traditional reliability statistics for criterion-referenced tests.
Construct validity	When a test is purported to measure a construct (i.e., a trait, intellectual process, or other unobservable characteristic of test takers), and it does so, it has construct validity.
Content validity	The term used to describe the efficacy of a test which measures the content or subject matter it is intended to measure. This type of validity is usually confirmed by the judgment of subject matter specialists who examine the test specifications and test items (cf. *descriptive validity*).
Correlation	The degree and direction of linear relation between two variables. Positive correlations describe direct relations and negative ones describe inverse relations. The degree of relation increases as the numerical value of the correlation statistic departs 0 and approaches +1.0 or -1.0.
Correlation coefficient	The number that describes a correlation.
Criterion	[a] In this volume and in many writings on CRT, the pool of potential test items measuring the same skill, objective, or attitude from which

the actual items on a CRT are a sample. The larger set of possible items which the given items represent.

[b] In other contexts, the cutting score or passing score. This general meaning of the term as a synonym for *standard* is misleading because standards may be expressed in absolute terms (e.g., 80% correct) or in comparative terms (e.g., 80th percentile). The latter is a norm-referenced standard, not a criterion-referenced one.

[c] The "real world" behavior or state which some types of test are designed to reflect. For example, a multiple choice test of composition skills is intended to identify the test taker's skills on the criterion of actual writing. A college entrance exam is meant to identify the future criterion of success in college.

Criterion group — A well defined group of test takers whose typical score serves as a standard of comparison for other students' scores. For example, state science fair medalists might be criterion groups whose typical scores on a science test could serve as a standard of high achievement. A random sample of students from the population for which a test is intended would be a criterion group whose typical scores could serve as a standard of average performance.

Criterion-referenced test (CRT) — A test designed so that the test items are "referenced to," or measure, the specific behaviors described in the criterion. The items for CRTs are supposed to be a representative sample of the criterion. CRTs are intended to show the extent to which a student possesses a particular skill or attitude.

Criterion validity — A high correlation of scores on a test with a criterion (cf. *criterion*[c]) for which the test is supposed to be an indicator. Predictive and concurrent validity are types of criterion validity.

Curriculum cross-reference — An index in which the items of a test are keyed to the pages or sections in published instructional materials that cover the same skills.

Cutting score	The score which serves as a dividing line between categories of achievement such as mastery and non-mastery or passing and not passing.
Decision rules	The rules for interpreting test scores in terms of categories of achievement, such as mastery/uncertain/non-mastery.
Description	See *test specifications.*
Descriptive validity	A term used to describe the efficacy of a test whose items accurately reflect the content, behavior, and format called for in its specifications. The specifications are then a valid description of the items or tasks.
Diagnostic test	A test that is designed to give information about a test taker's specific strengths and weaknesses within a subject area.
Discriminating power	The degree to which a test item distinguishes test takers who get high total scores on the test from those who get low total scores. Items are selected for norm-referenced tests so as to have high discriminating power.
Divergent validity	The validity that a test has when it measures the intended skill or attitude without being much affected by other, irrelevant skills, attitudes, or factors. For example, a math test lacks divergent validity if students' scores are greatly affected by the reading level of word problems. A test of reading comprehension lacks divergent validity if its scores are heavily influenced by pupils' general factual knowledge.
Domain	[a] the population of possible test items or tasks from which actual test items are sampled (cf. criterion[a]). [b] In other contexts, such as "cognitive domain" or "reading domain," the term refers to the general curricular area.
Domain-referenced test	A test that is designed so that test items are "referenced to," or measure, an individual's mastery of the population of tasks in a domain. Such a test yields information about the

	proportion of tasks within the domain that the test taker has mastered.
Domain specification	A form of test specification which describes in detail the characteristics of the total pool of potential items for measuring a specific skill or attitude. It is a technical document that deals with details of test content construction such as characteristics of distractors, rules for scoring, and rules for sampling items from the domain.
Factor analysis	A variety of statistical methods for identifying the distinct factors (e.g., abilities or traits or interests) that are reliably measured in a set of tests given to the same test takers.
False negative	The error of deciding that a student does not have mastery knowledge when (s)he actually does, i.e., failing to pass a deserving student.
False positive	The error of deciding that a student has mastery knowledge when (s)he actually does not, i.e., passing an undeserving student.
Field test	A tryout of a test in the actual conditions under which it will be used. Information from such tryouts is used to improve the test and establish norms and validity.
Grade equivalent score	A form of derived score for NRTs which is supposed to tell, for any raw score, the grade level, in years and months, for which that raw score is the national average. Owing to the misleading nature of grade equivalent scores, the professional test standards (APA, 1974) discourage the use of grade equivalents.
Individualization	Designing instruction to meet the particular needs of the individual student. Criterion-referenced measurement is useful for individualizing because it facilitates identification, by objective, of individual students' strengths and needs.
Inter-item correlation	The correlation among items on the same test, taken to show the degree to which the items are measuring the same thing.

Item	An individual task or question on a test.
Item analysis	The process of looking at students' scores on test items to determine such things as the items' difficulty levels and consistency in discriminating between high and low scorers. Items are analyzed for the purpose of identifying those which are good and those which are poor.
Item by group interaction	The case where the items on a test which are hardest for one group of test takers (e.g., one race or one gender) are different from the items which another group finds hardest. A form of evidence for bias in the test.
Item form	A type of test specification which states in complete detail the properties of items on a test. It does so by laying out a frame of text which is to be constant for all of the test items, then specifying the variable values that may go into specified slots in the frame, and rules for selecting among the possible variable values. An item form includes the instructions or additional information given to the test taker and describes the appropriate answer method. It also defines the correct responses.
Item generation	The process of constructing test tasks, items, or questions.
Item-objective congruence	The type of validity based on evidence that a test's items are consistent with its specifications.
Item uniformity	The characteristic a test exhibits when all test items measure a uniform, coherent skill or attitude (when the skill or attitude itself is uniform). Item uniformity is determined by factor analysis, inter-item correlations, and item-test correlations.
Level	Age or grade placement for which a test is designed.
Mastery score	The score on a particular test which indicates that a test taker has reached a predetermined level of proficiency.

Mastery test	A test designed to determine the extent to which test takers have learned or become proficient in a given unit, concept, topic, or skill.
Norms	One type of comparative information for interpreting norm-referenced tests. Norms are usually given in the form of percentiles. They describe the ranking of each possible score among the students who were in the test's field tryouts, but do not indicate the absolute degree of skill or mastery that is exhibited by the scores.
Norm-referenced test	In achievement testing, a test that is designed to survey the skills and knowledge common to most educational programs. This type of test yields information about how individual test takers' scores compare with the scores of the others who have also taken the test and provides only a very general description of the skills or attitudes being measured.
Objectives-based (or objectives-referenced) test	A test designed so that the items assess specified objectives for the purpose of making a mastery/non-mastery decision about the test taker.
Percentile	A number which indicates the percentage of scores which fall below a given test score. For example, a test taker in the 95th percentile scored higher than 95% of the students in the norm group. Small differences in raw scores sometimes make large differences in percentile ranking, especially in the middle percentiles. Percentiles thus should not be taken as a direct or absolute measure of learning.
Practice effect	The change in a test taker's score that is due to previous experience with the same or similar test rather than to a change in the skill or attitude to be measured.
Prescriptive	Suggesting materials or activities for teaching and learning particular skills.
Program-embedded test	A test that either is not sold apart from a body of curricular materials or that refers so

	closely to specific curricular materials that it would be unsuitable for testing students who had used other texts, practice exercises, etc.
r	A sumbol that stands for correlation coeffi-cient.
Random sample	A sample that is drawn from the total population (of students or schools or test items) so that every member of the population has an equal chance of being selected. This procedure is used to avoid bias in selecting the sample.
Reference group	A well defined group whose scores are used as a standard of comparison.
Reliability	The stability or consistency with which a test measures a skill or attitude. Absence of incidental fluctuations in score. Several types of reliability are distinguished: consistency of individuals' scores from one occasion to another (test-retest); consistency from one form of a test to another (alternate forms); and consistency among the items themselves (internal consistency or split half). Either the total test scores or the instructional decisions based on the test scores may be studied for their reliability.
Response materials	The materials a test taker uses for recording answers to a test (e.g., test booklets, answer sheets).
Response mode	The answer form a test requires (e.g., multiple choice, true-false, short answer, essay).
Response spaces	Places provided on a test form or answer sheet for recording answers.
Sample item	A sample test question given as part of the instructions to students to show them how to take the test.
Sampling plan (sampling rule)	The selection procedure that is followed to ensure that a sample represents the total group from which it was drawn.
Sensitivity to learning	A test's ability to detect an increase in the test taker's knowledge or skill.

Social fairness	The quality a test exhibits when test content does not stereotype or disparage any social group (i.e., any race, language group, gender, etc.).
Specifications	See *test specifications*.
Specimen set	A collection of test materials that serve as a sample of the complete test package. Many publishers sell these materials to enable test users to decide whether to buy the entire testing system.
Standard	[a] A degree or amount of quality, excellence, or attainment. [b] A basis of comparison.
Standardized test	[a] A norm-referenced test. [b] A test that has been designed so that all testees take the test under similar conditions. This latter usage may lead to some confusion since it may include criterion-referenced tests, unlike meaning [a].
Statistically significant difference	A difference in scores or numbers that is large enough as to be unlikely a result of mere chance.
Stem	The question or stimulus part of a test item as opposed to the response choices or responses.
Stimulus	The item stem and any other information, such as a graph or picture, that is used to pose the question in a test item and to elicit the response.
Stratified random sample	A sample made by first dividing a population (of people or test items) into naturally occurring groups (strata), then sampling from each in proportion to its relative size.
Template	A scoring overlay with the pattern of correct answers perforated to facilitate hand scoring.
Test	A tool for finding out how well students know a body of information, have mastery of a skill,

	or possess an attitude. The tool involves presenting some stimuli (or questions) to elicit responses from the students. Checklists and observation schedules are not considered tests in this context.
Test specifications	The description of the set of possible items for a test and directions for sampling items from that set. This description tells what is to be measured, and how. It serves as directions to the test writer for constructing a test.
Validity	A test or measure has validity if its scores mean what they are supposed to mean. There are different types of validity (cf. *content*, *descriptive*, *criterion*, and *construct validity*), each one verified in a somewhat different way.

APPENDIX D
Supplement to Chapter 3: Example of a Domain Description Which Would Receive a Level A Rating

Domain Title

Applying principles of U.S. foreign policy[1]

General Description

Given a description of a fictional international situation in which the United States may wish to act and the name of American foreign policy document or pronouncement, the students will select from a list of choices the course of action that would most likely follow from the given document or pronouncement.

Sample Item

Directions: Below are some made-up stories about world events. Answer each question by picking a choice and writing its letter on the answer sheet.

Some Russian agents became members of the Christian Democratic Party in Chile. The party attacked the President's house and arrested him. The Russian agents set themselves up as President and Vice-President of Chile. Chile then asked to become an "affiliated republic" of the U.S.S.R.

Based on the Monroe Doctrine, what would the U.S. do?

a. Ignore the new status of Chile.
b. Warn Russia that its influence is to be withdrawn from Chile.
c. Refuse to recognize the new government of Chile because it came to power illegally.
d. Send arms to all groups in the country that swear to oppose communism.

---

[1]This domain by Clinton B. Walker is reprinted from *Illustrative Test Specifications for the USDESEA Matrix of Educational Objectives*. W. J. Popham, Project Director. Los Angeles: Educational Objectives and Measures, 1976.

Stimulus Attributes

1. The fictional passage will consist of 50 words or less followed by the name of a foreign policy pronouncement or document inserted into the question, "Based on the _____, what will the U.S. do?

2. The policy named in the stimulus passage will be a document or pronouncement selected from the Domain Supplement.

3. Each passage will consist of two parts: a) a background description of an action taken by a foreign nation, and b) a statement of the action to which the foreign policy document or pronouncement is to be applied.

   a. The background statement will be analogous to an historical situation which either preceded the document or pronouncement, or for which the document or pronouncement was used. For example, the Monroe Doctrine was laid down in response to European designs on American nations that were attempting to establish independence. A parallel case today might describe a European country trying to encroach on the sovereignty of such a country.

   b. The statement of an action will be an action taken by a real foreign nation that conforms to one of the following categories:

      1. Initiation of an international conflict.
      2. Initiation of a civil conflict. This may include coups, revolutions, riots, protest marches, civil war, or a parliamentary crisis.
      3. Initiation of an international relationship. This includes trade negotiations, friendship pacts, military alliances, and all classes of treaties.
      4. Appeal for foreign aid to meet economic or military needs.
      5. Development and stockpiling of military weapons.

4. All statements in the passage will refer to specific nations and events. Descriptions such as, "A nation is at war with another country," are not acceptable. The events described may be set in the present or past, as appropriate.

5. When the document or pronouncement mentioned in the stimulus passage is tied to a particular geographic region, countries named in the passage must belong to that region.

6. Passages will be written no higher than the 8th grade reading level.

Response Attributes

1. Students will mark the letter of one of the four given response alternatives.

2. The correct response will be a course of action that is governed by the main principles of the document or pronouncement named in the stem.

3. Response choices consist of the correct response and three distractors. Each choice will have the following characteristics:

   a. Describe a specific course of action that refers to the people, nations, and actions in the stimulus passage.

   b. Be brief phrases written to complete the understood subject, "The United States would . . ."

4. Distractors will be written to meet these additional criteria:

   a. At least one distractor will describe an action derived from a different document or pronouncement selected from the Domain Supplement

   b. Distractors will be plausible courses of action, not fanciful.

Domain Supplement

Foreign Policy Documents and Pronouncements:

The following list of foreign policy pronouncements and documents was selected from Brockway, T., *Basic Documents in United States Foreign Policy*. Princeton, NJ: D. Van Nostrand Company, 1968. The documents were chosen on the basis of their historical impact or potential application to current events. The list appears in chronological order.

1. Washington's Farewell Address
2. The Monroe Doctrine
3. Webster on Revolutions Abroad
4. Open Door in China
5. The Platt Amendment
6. Roosevelt Corollary of the Monroe Doctrine
7. The Fourteen Points
8. The Washington Conference
9. The Japanese Exclusion Act
10. The Kellogg-Briand Pact
11. The Stimson Doctrine
12. Roosevelt's Quarantine Speech
13. The Atlantic Charter
14. The Connally Resolution
15. The Yalta Agreements
16. The Potsdam Agreement
17. United States Proposals for the International Control of Atomic Power
18. The Truman Doctrine
19. The Marshall Plan
20. The Point Four Program
21. The North Atlantic Treaty
22. American-Japanese Defense Pact
23. Atoms for Peace: Eisenhower's Proposal to the United Nations
24. The Eisenhower Doctrine
25. Alliance for Progress
26. Kennedy's Grand Design
27. Treaty on the Peaceful Uses of Outer Space

APPENDIX E
Available Tests That Were Screened Out of the Pool of Measures Reviewed in This Volume

CRTs that are embedded in a specific curriculum

Clues to Reading Progress
Educational Progress Corporation

Communication Skills Program
Ginn & Company

Competency Skills Test for Keys to Reading
Economy Company

Competency Skills Tests for Keys to Independence in Reading
Economy Company

Continuous Progress Laboratories
Educational Progress Corporation

Criterion Assessment Tests
J. B. Lippincott

Dale Avenue Project
Paterson (NJ) School District

Developing Mathematical Processes
Rand McNally

Developmental Syntax Program
Learning Concepts

Gaining Math Skills
McCormick-Mathers Publishing Company

Holt Basic Reading System
Holt, Rinehart, Winston

Individualized Mathematics Program
Educational & Industrial Testing Service

Individualized Mathematics System
Ginn & Company

Individualized Science Program CRTs
Imperial International Learning Corporation

Learning Staircase
Learning Concepts

Math Management System Placement Test
Clark County School District

Mathematics Around Us
Scott Foresman & Company

Mathematics Laboratory
McCormick-Mathers Publishing Company

Perceptual Skills Curriculum
Walker Educational Book Company

Progressive Achievement Tests
New Zealand Council for Educational Research

Project ACTIVE CRTs
ACTIVE, Ocean Township (NJ)
Elementary School

Series m: Macmillan Math
Macmillan Company

Series r: Macmillan Reading
Macmillan Company

System 80
Borg-Warner Educational Systems

System for Teacher Evaluation
of Prereading Skills
CTB/McGraw-Hill

Teaching Essential Language &
Reading
Educational & Industrial
Testing Service

SWRL Kindergarten Program
Ginn & Company

Tests received in response to our search, but screened out of the pool of tests to be reviewed*

Listed below are the names and publishers of tests which were screened out. The reasons for excluding each are given, keyed to the following list:

1. The test became unavailable before publication of this volume.
2. The skills measured are a usual result of maturation or general experience.
3. The test is not built around explicit objectives.
4. Items are not keyed to objectives.
5. There is only one item per objective.
6. Scores for the separate objectives are not given.
7. Scores are not interpreted in terms of proficiency or mastery.
8. The test was not designed as an objectives-based measure.
9. The test was not available to review in time for inclusion in this volume.

---

*No judgment about the merits of these tests is intended by their being excluded. Tests not meeting criteria 3 through 8 are not CRTs.

ACER Class Achievement Tests in Mathematics (3,8)
Australian Council for Educational Research

APPEL Test (1)
Insgroup (formerly EDCODYNE)

Assessment of Career Development (8)
American College Testing Program

Basic School Skills Inventory (5,7)
Follett Publishing Company

Boehm Test of Basic Concepts (3 or 5, 8)
Psychological Corporation

Brigance Diagnostic Inventory of Basic Skills (9)
Walker Educational Book Corporation

Cincinnati Mathematics Inventory (5)
Cincinnati Public Schools, Dept. of Research & Development

Composite Auditory Perception Test (8)
Alameda County (CA) School Dept.

Criterion-Referenced Tests for Reading and Writing in A Technology of Reading and Writing, Vol. 2 (9)
Academic Press

Delco Readiness Test (3,7)
Walter M. Rhoades

Development Test of Visual Motor Integration (2,8)
Follett Publishing Company

Diagnostic Skills Battery (9)
Scholastic Testing Service

Emporia State Algebra II Test (5,8)
Bureau of Educational Measurements, Emporia Kansas State College

Individual Phonics Criterion Test (5,7)
Dreier Educational Systems

Kraner Preschool Math Inventory (3)
Learning Concepts

NM Attitude Toward Work Test (7)
Monitor

Oral Reading Criterion Test (3,8)
Dreier Educational Systems

PIRAMID (1)
PIRAMID Consortium

Preschool Attainment Record (5)
American Guidance Service

Prescriptive Mathematics Inventory (1)
CTB/McGraw-Hill

Reading Management System, Diagnostic Step Tests (4,5,6)
Clark County School District

Reading Skills Survey Tests (5,6)
Economy Company

Self-Directed, Interpretative and Creative Reading (4,6)

Senior High Assessment of Reading Performance (SHARP) (9)
CTB/McGraw-Hill

SRA Reading Record (8)
Science Research Associates

Stanford Achievement Test (8)
Harcourt Brace Jovanovich

Stanford Test of Academic Skills (TASK) (8)
Harcourt Brace Jovanovich

Visual Analysis Test (8)
University of Pittsburgh

INDEX A
Names of Reviewed Tests

INDEX B
Tests by Subject Matter

MATHEMATICS INDEX

UNDERSTANDING MATH CONCEPTS: numbers and sets; numeral systems and number principles; number relationships; and ordering numbers and symbols

PERFORMING ARITHMETIC OPERATIONS: whole number computations - addition, subtraction, multiplication, division

PERFORMING ARITHMETIC OPERATIONS: fractions, decimals, and percentage computations - addition, subtraction, multiplication, division

APPLYING MATHEMATICS: problem solving, word problems

GEOMETRY OPERATIONS AND RELATIONS

MEASUREMENT: weight, volume, length, angular, time, speed

USE OF TABLES, GRAPHS, STATISTICAL CONCEPTS

READING INDEX

VOCABULARY/WORD RECOGNITION: auditory and visual

READING COMPREHENSION: literal meaning (main idea)

READING COMPREHENSION: interpretative meaning

WRITING SKILLS: spelling, punctuation, and grammatical skills

REFERENCE STUDY SKILLS AND TECHNIQUES

APPRECIATION OF READING (dictionaries, newspapers, books)

OTHER SUBJECTS INDEX

INDEX C
Publishers' Names and Addresses

Publisher	Tests	Page
Academic Therapy Publications 1539 Fourth Street P.O. Box 899 San Rafael, CA 94901	Criterion Test of Basic Skills	49
American Guidance Service (AGS) Publishers' Building Circle Pines, MN 55014	Doren Diagnostic Reading Test of Word Recognition Skills	58
	KeyMath Diagnostic Arithmetic Test	80
	Woodcock Reading Mastery Tests	152
Cal Press, Inc. 76 Madison Avenue New York, NY 10016	REAL: Reading/Everyday Activities in Life	126
Cooperative Educational Service Agency #13 908 W. Main Street Waupun, WI 53963	Early Childhood Assessment	60
Croft Educational Services 4922 Harford Road Baltimore, MD 21214	Cooper-McGuire Diagnostic Word Analysis Test	44
	McGuire-Bumpus Diagnostic Comprehension Test	110
CTB/McGraw-Hill Del Monte Research Park Monterey, CA 93940	Diagnostic Mathematics Inventory	56
	Everyday Skills Tests: Reading, Test A; Mathematics, Test A	62
	Prescriptive Reading Inventory	120
CTB/McGraw-Hill Ryerson Limited 330 Progress Avenue Scarborough, Ontario CANADA MIP 225	Pre-Reading Assessment Kit	118
Dallas Independent School District ATTN: Mr. Dean Arrasmith 3801 Herschel Street Dallas, TX 75219	Survey of Reading Skills	140

Publisher	Tests	Page
Dreier Educational Systems P.O. Box 1291 Highland Park, NJ 08904	Basic Word Vocabulary Test	38
	Group Phonics Analysis Test	68
	Instant Word Recognition Test	78
Educational and Industrial Testing Service (EdITS) P.O. Box 7234 San Diego, CA 92107	Tests of Achievement in Basic Skills - Math	142
	Tests of Achievement in Basic Skills - Reading and Language	144
Educational Progress Educational Development Corporation P.O. Box 45663 Tulsa, OK 74145	Individualized Criterion-Referenced Testing - Math	74
	Individualized Criterion-Referenced Testing - Reading	76
Educators Publishing Service 75 Moulton Street Cambridge, MA 02138	Sipay Word Analysis Tests	128
Follett Publishing Co. Department DM 1010 W. Washington Blvd. Chicago, IL 60607	Language and Thinking Program: Mastery Learning Criterion Tests	82
Harcourt Brace Jovanovich (see The Psychological Corporation)		
Houghton Mifflin 777 California Avenue Palo Alto, CA 94304	Individual Pupil Monitoring System - Mathematics	70
	Individual Pupil Monitoring System - Reading	72
Imperial International Learning Corp. (IIL) P.O. Box 548, Route 50 South Kankakee, IL 60901	Basic Arithmetic Skill Evaluation (BASE) and BASE II	36
Instructional Objectives Exchange (IOX) P.O. Box 24095 Los Angeles, CA 90025	Language Arts: Composition, Library, and Literary Skills	84
	Language Arts: Mechanics and Usage	86
	Language Arts: Word Forms and Syntax	88
	Mathematics: Elements, Symbolism, and Measurement	96

Publisher	Tests	Page
	Mathematics: Geometry	98
	Mathematics: Geometry, Operations, and Relations	100
	Mathematics: Measurement	102
	Mathematics: Numeration and Relations	104
	Mathematics: Operations and Properties	106
	Mathematics: Sets and Numbers	108
	Reading: Comprehension Skills	122
	Reading: Word Attack Skills	124
	Social Studies: American Government	132
J. B. Lippincott Company Educational Publishing Division East Washington Square Philadelphia, PA 19105	Beginning Assessment Test for Reading	40
Monitor P.O. Box 2337 Hollywood, CA 90028	New Mexico Career Education	112
	New Mexico Concepts of Ecology Test	114
	New Mexico Consumer Mathematics Test & Consumer Rights and Responsibilities Test	116
Multi-Media Associates, Inc. EPIC Criterion-Referenced Test Division P.O. Box 13052 4901 E. Fifth Street Tucson, AZ 85732	Criterion-Referenced Tests of Basic Reading and Computational Skills	46
NCS Educational Systems 4401 West 76th Street Minneapolis, MN 55435	Design for Math Skill Development	50
	Wisconsin Design for Reading Skill Development: Comprehension	146
	Wisconsin Design for Reading Development: Study Skills	148
	Wisconsin Design for Reading Skill Development: Word Attack	150

Publisher	Tests	Page
The Psychological Corporation A division of Harcourt Brace Jovanovich 757 Third Avenue New York, NY 10017	Skills Monitoring System: Reading	130
	Stanford Diagnostic Mathematics Test	136
	Stanford Diagnostic Reading Test	138
Revrac Publications Dr. Ronald P. Carver 10 W. Bridlespur Drive Kansas City, MO 64114	Carver-Darby Chunked Reading Test	42
Scholastic Testing Service, Inc. (STS) 480 Meyer Road Bensenville, IL 60106	Analysis of Skills (ASK) - Language Arts	30
	Analysis of Skills (ASK) - Mathematics	32
	Analysis of Skills (ASK) - Reading	34
Science Research Associates, Inc. (SRA) 259 East Erie Street Chicago, IL 60611	Diagnosis: An Instructional Aid - Mathematics	52
	Diagnosis: An Instructional Aid - Reading	54
	Mastery: An Evaluation Tool (Mathematics)	90
	Mastery: An Evaluation Tool (SOBAR Reading)	92
	SRA Survival Skills in Reading and Math	134
U-SAIL (Utah System Approach to Individualized Learning) 2971 Evergreen Avenue P.O. Box 9327 Salt Lake City, UT 84109	Math Diagnostic/Placement Tests	94
Richard L. Zweig Associates Testing Division 20800 Beach Blvd. P.O. Box 73 Huntington Beach, CA 92648	Fountain Valley Teacher Support System in Mathematics	64
	Fountain Valley Teacher Support System in Reading	66

REFERENCES

The references are listed under the following categories:

- Those cited in the text
- Those which provided lists of tests for review in this book
- Recommended reading

TEXTUAL

American Psychological Association (APA). *Standards for educational and psychological tests.* Washington, DC: APA, 1974.

Armbruster, B. B., Stevens, R. J., & Rosenshine, B. *Analyzing content coverage and emphasis: A study of three curricula and two tests.* Technical Report #26. Urbana, IL: Center for the Study of Reading, University of Illinois, 1977.

Baker, E. L. *Achievement testing in urban schools: New numbers.* To be published by CEMREL, Inc., St. Louis, Missouri, in the Urban Education Monograph Series, Margaret Solomon (Ed.).

Barta, M. B., Ahn, J. R., & Gastright, J. F. Some problems in interpreting criterion-referenced test results in a program evaluation. *Studies in Educational Evaluation*, 1976, *2*(3), 193-202.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, *56*, 81-105.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental design for research.* Chicago: Rand McNally, 1963.

Cronbach, L. J. *Essentials of psychological testing* (3rd ed.). New York: Harper & Row, 1970.

Denham, C. H. *Score reporting and item selection in selected criterion referenced and domain referenced tests.* Paper given at the annual meeting of the National Council on Measurement in Education, New York, April, 1977.

Dotseth, M., Hunter, R., & Walker, C. B. *Survey of test selectors' concerns and the test selection process.* CSE Report #107. Los Angeles: Center for the Study of Evaluation, University of California, 1978.

Ebel, R. L. *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall, 1972.

Floden, R. E., Porter, A. C., Schmidt, W. H., & Freeman, D. J. Don't they all measure the same thing? Consequences of standardized test selection. In E. L. Baker & E. S. Quellmalz (Eds.), *Educational testing and evaluation: Design, analysis, and policy.* Beverly Hills, CA: Sage Publications, 1979.

Guion, R. M. Content validity--the source of my discontent. *Applied Psychological Measurement*, 1977, *1*, 1-10.

Hambleton, R., & Eignor, D. *Guidelines for evaluating criterion-referenced tests and test manuals.* Paper delivered at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of the technical issues and development. *Review of Educational Research*, 1978, *48*, 1-47.

Hoepfner, R. Achievement test selection for program evaluation. In M. J. Wargo & D. R. Green (Eds.), *Achievement testing of disadvantaged and minority students for educational program evaluation.* Monterey, CA: CTB/McGraw-Hill, 1978.

Hoefpner, R., et al. *CSE secondary school test evaluations.* Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Hoepfner, R., et al. *CSE elementary school test evaluations.* Los Angeles: Center for the Study of Evaluation, University of California, 1976

Jenkins, J. R., & Pany, D. *Curriculum biases in reading achievement tests.* Technical Report #16. Urbana, IL: Center for the Study of Reading, University of Illinois, 1976

Katz, M. *Selecting an achievement test.* Princeton, NJ: Educational Testing Service, 1973.

Linn, R. L. *Issues of validity in measurement for competency-based programs.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, April, 1977.

Linn, R. L., & Slinde, J. A. The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research*, 1977, *47*(1), 121-150.

Lyon, C. D., Doscher, L., McGranahan, P., & Williams, R. *Evaluation and school districts.* Los Angeles: Center for the Study of Evaluation, University of California, 1978.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, *30*, 955-966.

Popham, W. J. *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

Shoemaker, D. M. Evaluating the effectiveness of competing instructional programs. *Educational Researcher*, 1972, *1* (May), 5-12.

Stake, R. E. *More subjective!* Remarks made in an invited debate on the question, "Should educational evaluation be more objective or more subjective?" at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Stallard, C. Comparing objective based reading programs. *Journal of Reading*, 1977, *21*(1), 36-44.

Stallard, C. Managing reading instruction: Comparative analysis of objective-based reading programs. *Educational Technology*, 1977, *17*(12), 21-26.

Tallmadge, G. K., & Horst, D. P. *A procedural guide for validating achievement gains in educational projects.* Washington, DC: Government Printing Office, 1976. (GPO Stock Number 017-080-01516-1.)

Tripodi, T., Fellin, P., & Epstein, I. *Differential social program evaluation.* Itasca, IL: Peacock, 1978.

Walker, C. B. *Control test items: A baseline measure for evaluating achievement.* Paper presented at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Walker, D. F., & Schaffarzik, J. Comparing curricula. *Review of Educational Research*, 1974, *44*, 83-112.

TEST LISTS

Barrett, J. E. (Ed.). *Where behavioral objectives exist.* Norton, MA: Project SPOKE, 1974.

Education programs that work. San Francisco: Far West Regional Laboratory for Educational Research and Development, 1975.

Gitlin, C. *Review of commercially available criterion-referenced tests.* Final Report, Contract No. DAJA37-75-C-1760. United States Dependents Schools, European Area. Los Angeles: Educational Objectives and Measures, February, 1976.

Keller, C. M. *Criterion-referenced measures: A bibliography.* Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1972. (ED 060 041, TM 001 124.)

Knapp, J. *A collection of criterion-referenced tests.* TM Report No. 31. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation, December, 1974. (ED 099 427.)

Rosen, P. (Ed.). *Test collection bibliographies: Criterion-referenced measures.* Princeton, NJ: Educational Testing Service, 1973. (ED 104 910, TM 004 362.) (Includes supplement dated August, 1974.)

Rosen, P. (Ed.). *Test collection bulletin.* Princeton, NJ: Educational Testing Service, 1975 (Vol. 9), and 1976 (Vol. 10).

Test library catalog (revised edition). Los Angeles: Los Angeles County Superintendent of Schools, Division of Program Evaluation, Research and Pupil Services, 1976.

RECOMMENDED READING

Airasian, P. W., & Madaus, G. F. Criterion-referenced testing in the classroom. *Measurement and Education*, 1972, *3*, 73-88.

Baker, E. L. Cooperation and the state of the world in criterion-referenced tests. *Educational Horizons*, 1974, *52*(4), 193-196.

Block, J. H. Criterion-referenced measurement: Potential. *School Review*, 1971, *79*, 289-297.

Boehm, A. E. Criterion-referenced assessment for the teacher. *Teachers College Record*, 1973, *75*(1), 117-126.

Carver, R. P. Two dimensions of tests: Psychometric and edumetric. *American Psychologist*, 1974, July, 512-578.

Ebel, R. L. Criterion-referenced measurements: Limitations. *School Review*, 1971, *79*, 282-288.

Ebel, R. L. Criterion-referenced and norm-referenced measurements. In *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall, 1972, 83-86.

Ebel, R. L. Evaluation and educational objectives. *Journal of Educational Measurement*, 1973, *10*(4), 273-279.

Esler, W. K., & Dziuban, C. D. Criterion referenced test, some advantages and disadvantages for science instruction. *Science Education*, 1974, *58*(2), 171-174.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 1963, *18*, 519-521.

Glass, G. V. *Standards and criteria*. San Mateo, CA: San Mateo Educational Resource Center, 1977. (No. ID 005 555, 55 pages)

Good, T. L., Biddle, B. J., & Brophy, J. E. Criterion-referenced testing. In *Teachers make a difference*. New York: Holt, Rinehart & Winston, 1975.

Gronlund, N. E. *Preparing criterion-referenced tests for classroom instruction*. New York: Macmillan Company, 1973.

Haladyna, T. *The paradox of criterion-referenced measurement*. (ERIC Number ED 126 155, April, 1976, 25 pages.)

Harsh, J. R. *The forests, trees, branches and leaves, revisited: Norm, domain, objective and criterion-referenced assessments for educational assessment and evaluation.* Association for Measurement and Evaluation in Guidance Monograph No. 1. Los Angeles: California Personnel and Guidance Association, February, 1974.

Hively, W. *Domain referenced testing.* Englewood Cliffs, NJ: Educational Technology Publications, 1974.

Hively, W. Introduction to domain-referenced testing. *Educational Technology*, 1974, *14*(6), 5-10.

Hocker, R., Green, D. R., Ginsburg, N., & Hyman, H. *The nature and uses of criterion-referenced and norm-referenced achievement tests.* Special Report, Vol. 4, No. 3. Burlingame, CA: Association of California School Administrators, undated (probably 1975).

Martuza, V. R. *Applying norm-referenced and criterion-referenced measurement in education.* Boston: Allyn and Bacon, 1977.

Mehrens, W. A., & Lehmann, I. J. Norm- and criterion-referenced measurement. In *Measurement and evaluation in education and psychology.* New York: Holt, Rinehart & Winston, 1973, 63-76.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications.* Berkeley, CA: McCutchan, 1974. (Also available as a separate monograph.)

Millman, J. Program assessment, criterion-referenced tests, and things like that. *Educational Horizons*, 1974, *52*(4), 188-192.

Popham, W. J. *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

Popham, W. J. (Ed.). *Criterion-referenced measurement: An introduction.* Englewood Cliffs, NJ: Educational Technology Publications, 1971.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, *6*, 1-9.

Sanders, J. R., & Murray, S. L. Alternatives for achievement testing. *Educational Technology*, 1976, *16*(3), 7 pages.

285

KEY TO THE EVALUATIVE SECTIONS OF CSE TEST REVIEWS*

MEASUREMENT PROPERTIES: CONCEPTUAL VALIDITY

1. Domain Descriptions. How good (i.e., thorough and comprehensive) are the descriptions of the objectives or domains to be tested?
   - A. Very good (objectives are thoroughly described)
   - B. Adequate (objectives are stated behaviorally but not in detail)
   - C. Poor (objectives are loosely described and subject to various interpretations)
2. Agreement. How well do the test items match their objectives?
   - A. The match is confirmed by sound evidence
   - C. Data are not provided or are not persuasive
3. Representativeness. How adequately do the items sample their objectives?
   - A. Items are representative of domains
   - C. Item selection is either unrepresentative or unreported

MEASUREMENT PROPERTIES: FIELD TEST VALIDITY

4. Sensitivity. Does conventional instruction lead to test-score gains?
   - A. Test scores reflect instruction
   - C. Data are not provided or are not persuasive
5. Item Uniformity. How similar are the scores on the different items for an objective?
   - A. Some evidence of item uniformity is provided
   - C. No data are provided
6. Divergent Validity. Are the scores for each objective relatively uninfluenced by other skills?
   - A. Independence of skills is confirmed
   - C. Data are not provided or are not persuasive
7. Lack of Bias. Are test scores unfairly affected by social group factors?
   - A. Persuasive evidence of lack of bias is offered for at least two groups (e.g., women, specific ethnic groups
   - C. Data are not provided or are not persuasive
8. Consistency of Scores. Are scores on individual objectives consistent over time or over parallel test forms?
   - A. Consistency of scores for objectives is shown over parallel forms or repeated testing
   - C. Data are not provided

APPROPRIATENESS AND USABILITY

9. Clarity of Instructions. How clear and complete are the instructions to students?
   - A. Instructions are clear, complete, and include sample items
   - B. Either instructions or sample items are lacking
   - C. Both are lacking
10. Item Review. Does the publisher report that items were either logically reviewed or field tested for quality?
    - A. Yes
    - C. No
11. Visible Characteristics. Is the layout and print easily readable?
    - A. Print and layout are readable for more than 90% of objectives
    - C. At least 10% of objectives have problems in readability
12. Ease of Responding. Is the format for recording answers appropriate for the intended students?
    - A. Responding is easy for more than 90% of the objectives
    - C. Lack of clarity, crowding, etc., make responding difficult in at least 10% of objectives
13. Informativeness. Does the test buyer have adequate information about the test before buying it?
    - A. Yes
    - C. No
14. Curriculum Cross-Referencing. Are the test objectives indexed to at least two series of relevant teaching materials?
    - A. Yes
    - C. No
15. Flexibility. Are many of the objectives tested at more than one level, and are single objectives easy to test separately?
    - A. Objectives are varied, carry over across test levels, and are easy to test separately
    - B. One feature is missing from variety, carry over, or separability
    - C. Two or three of the features are missing
16. Alternate Forms. Are parallel forms available for each test?
    - A. Yes
    - C. No
17. Test Administration. Are the directions to the examiner clear, complete, and easy to use?
    - A. Directions are clear, complete, and easy to use
    - C. One or more of the above features are missing
18. Scoring. Are both machine scoring and easy hand scoring available?
    - A. Yes
    - B. Easy, objective hand scoring is available, but no machine scoring
    - C. Hand scoring is not easy or objective; or only machine scoring is offered
19. Record Keeping. Does the publisher provide record forms that are keyed to test objectives and are easy to use?
    - A. Yes
    - C. They are not included or not keyed to test objectives
20. Decision Rules. Are well justified, easy-to-use rules given for making instructional decisions on the basis of test results?
    - A. Yes
    - C. Decision rules either are not given, not easy to use, or not justified
21. Comparative Data. Are scores of a representative reference group of students given for comparing with scores of pupils in the test user's program?
    - A. National norms, criterion group data, or item difficulty values are provided
    - C. These are not provided or are not clearly representative

*This system for evaluating CRTs is explained in detail in the text. For test features where only two levels of quality are distinguished, the letters A and C are used to indicate the levels.