DOCUMENT RESUME

ED 223 703 TM 820 850

AUTHOR Baker, Eva L.; And Others
TITLE Making, Choosing, and Using Tests: A Practicum on Domain-Referenced Testing.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Aug 80
CONTRACT NIE-78-0213
NOTE 205p.
PUB TYPE Guides - Non-Classroom Use (055)

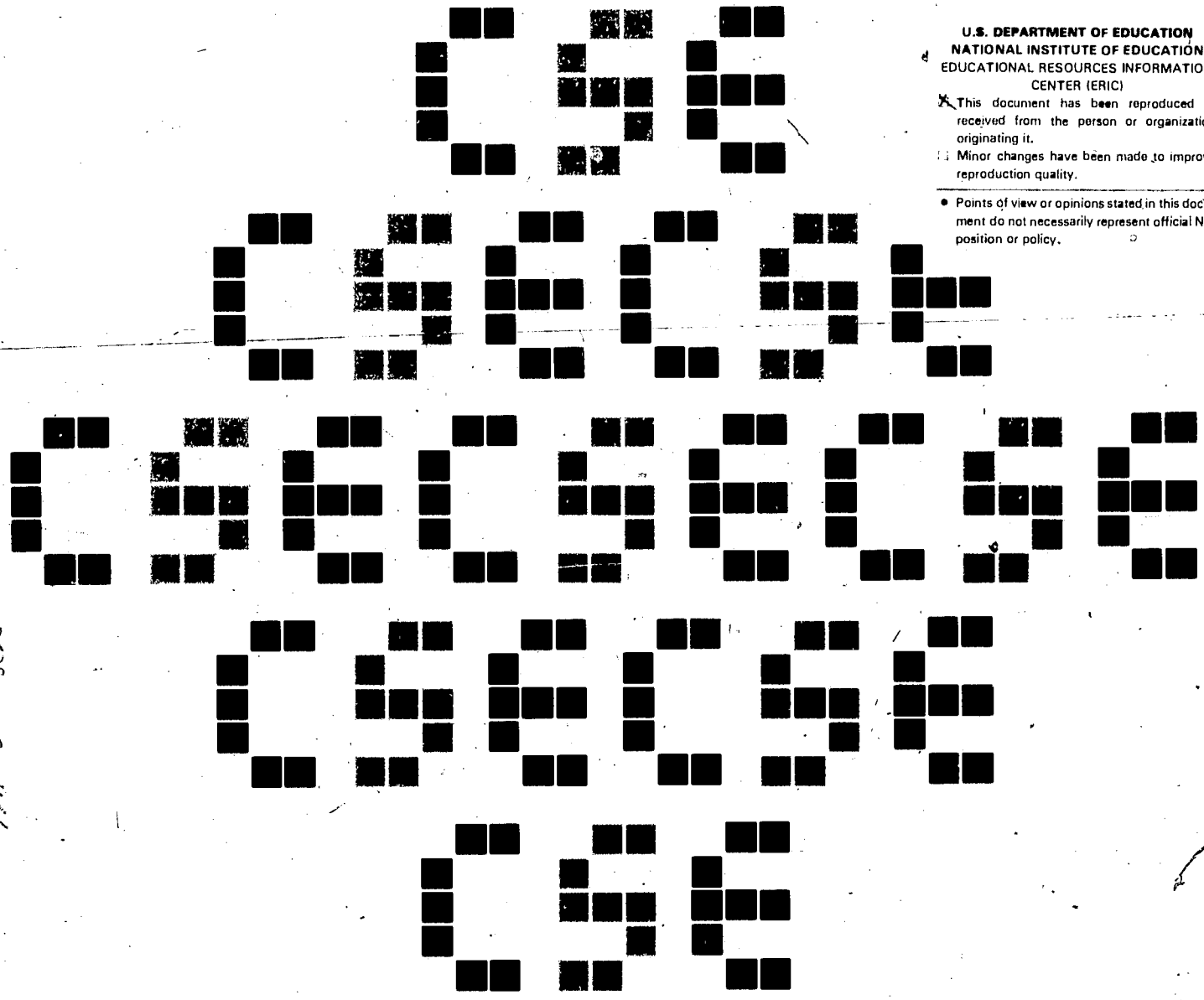
EDRS PRICE MF01/PC09 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Elementary Secondary Education; *Evaluation Criteria; Item Analysis; *Local Norms; Measurement Objectives; Relevance (Education); Skill Analysis; *Test Construction; Testing Programs; Test Items; Test Reliability; *Test Selection; Test Use; *Test Validity
IDENTIFIERS Domain Referenced Tests; Test Manuals

ABSTRACT

The materials presented were developed for use in a series of conferences on testing and instruction sponsored by the National Institute of Education, with the United States Office of Education, the UCLA Center for the Study of Evaluation, and a network of research and development agencies. They are intended for use by school practitioners and others concerned with the development or selection of tests geared toward local curricula and objectives. The development and validation process is described. The volume provides procedures for selecting or developing tests that are instructionally relevant and technically sound. Two procedures for test development rely on domain specifications and item review for congruence with these specifications. The two procedures for test selection are concerned with a test's relevance and its technical properties. Domain specifications connect learning outcomes to instructional content and the assessment of learning by providing rules for describing the domain, generating items, and setting their linguistic and cognitive complexity. The test selection procedures consider a test's instructional relevance to specified skills and objectives and its technical qualities. A training unit and practice materials for each procedure and a facilitator's guide are provided. (Author/CM)

Center for the Study of Evaluation

UCLA Graduate School of Education
Los Angeles, California 90024



Center for the Study of Evaluation

The mission of the Center for the Study of Evaluation is to conduct inquiry, from a variety of perspectives, into the nature of educational programs and services. Our commitment to inquiry into the field of evaluation grows from the belief that school practices and the competencies and satisfactions of those who participate in the educational enterprise can benefit from information collected in accordance with social science methodologies. Activities of CSE involve study of the instruments and methodologies for collecting information as well as the sociopolitical contexts of educational decisions as a means of contributing to the long-range growth in effectiveness of public education.

Information about CSE and its publications may be obtained by writing to:

Director, Public Information
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

MAKING, CHOOSING AND USING TESTS:

A PRACTICUM ON DOMAIN-REFERENCED TESTING

Submitted to
National Institute of Education

Eva L. Baker, Linda G. Polin,
James Burry, and Clinton Walker

OB-NIE-78-0213
P4

Eva L. Baker
Principal Investigator

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

This project was supported in whole or in part by the National Institute of Education, Department of Health, Education and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

August, 1980

TABLE OF CONTENTS

ABSTRACT

The materials in this volume were developed for use in a series of conferences on testing and instruction sponsored by the National Institute of Education, in collaboration with the United States Office of Education, the UCLA Center for the Study of Evaluation, and a network of research and development agencies. They are intended for use by school practitioners and others concerned with the development or selection of tests geared toward local curricula and objectives. The materials were tried out and validated in a process that began before the conferences, continued during the conferences, and was completed via external review and final modification after the conferences.

The volume provides procedures for selecting or developing tests that are instructionally relevant and technically sound. It offers two procedures to be used in the test development process, and two to help make the test selection process more systematic. The procedures for test development rely on domain specifications and item review for congruence with these specifications. The procedures for test selection are concerned with a test's relevance and its technical properties.

Domain specifications connect learning outcomes to instructional content and the assessment of learning. They are developed so that the test maker will understand instructional intentions and so develop appropriate test items. Specifications provide rules for describing the domain, generating items, and setting their linguistic and cognitive complexity. Item review deals with how well test items reflect the content of the domain and follow the rules for item generation.

The test selection procedures consider a test's instructional relevance and its technical qualities. Instructional relevance is judged in terms of how well the test matches specified skills and objectives. Technical qualities are judged in terms of what the test measures, how it was developed, its appropriateness for the examinees, and the scores it reports and the interpretations they permit.

The volume contains a training unit and practice materials for each of these four procedures. Included with the training materials is a facilitator's guide used by the person designated to provide the training.

PREFACE

The materials in this book are intended to be used in the provision of training for making, choosing, and using tests. The history of their development and validation is as follows.

In the Spring of 1979, the National Institute of Education, in collaboration with the United States Office of Education, the UCLA Center for the Study of Evaluation (CSE), and members of a nation-wide network of research and development agencies, sponsored a national colloquy on the role of testing in the public schools. Eight regional conferences were held as a vehicle to share information among approximately 1200 participants. These people came from the community, parent groups, and the professions of teaching, educational research, policy, and administration, and represented local, state, and national interests. Of prime significance was the conference theme that testing could have an important role in improving the effectiveness of instruction, but that much remained to be understood about testing needs and problems.

Each conference involved presentations from national and regional figures in testing and instruction. The conferences also provided initial training opportunity in test development and test selection to acquaint participants further with some of the newer ideas in the field. The opportunity to provide this kind of training has led to the development of the materials contained in this book. The substance of the materials derived from earlier NIE funded work at CSE, and consisted of materials dealing with (1) test development including domain-referenced test specifications and item review procedures, and (2) test selection from the standpoints of a test's instructional relevance for a given curriculum, and its technical and practical merits.

Before the materials were presented at the regional conferences, they were tried out in two local field settings. These local try-outs led to initial materials revision and the development of a facilitator's guide to be followed by the person using the materials to provide training. The materials were subject to a second period of review on the basis of the first two or three regional conferences, after which further revisions were made. We felt, however, that the materials, given their potential for use by and effect on classroom teachers, should be subject to further review by external experts, after all eight conferences had been conducted.

External consultants reviewed the materials in terms of their methodological soundness, the relevance and accuracy of the domain specification examples they provide, and from the standpoint of classroom application -- are they potentially useful to teachers? are they sufficiently comprehensive to allow users to consider development of a wide variety of assessment devices? are exemplary materials relevant and accurate? In addition, both external reviewers examined the facilitator's guide accompanying the materials. Concurrent with these reviews, the materials were tried out once again with teaching and central office staff of a school district.

After the reviews and the trial, the materials were subject to a final level of scrutiny -- this time by CSE staff with a background in tests and testing and who represented classroom experience and content knowledge in the skills dealt with in the materials.

These CSE staff (1) independently of each other and of the external reviewers provided detailed critiques of the materials and made suggestions for revision; (2) examined the results of the critiques made by the

external reviewers, and (3) made appropriate changes to the training materials and the accompanying facilitator's guide. These changes are incorporated in the present materials.

The entire package of materials views the test development or selection process in terms of specified content areas and instructional strategy. In doing this, they begin the first step in linking the results of testing to instructional practice, thereby improving the effectiveness of instruction.

The materials in the book provide domain-referenced procedures for making, choosing, and using tests. The book offers two procedures to be used where there is reasonable control over the test development process; e.g., in the design of a district testing program or the development of teacher-made tests. It offers two procedures to be used for making test selection more systematic; e.g., in cases where the test is prepared by others such as a commercial test firm or a consultant. The procedures for test development (1) provide a blueprint for writing domain-referenced test specifications and (2) a technique for reviewing test items for congruence with the written specifications. The procedures for test selection are concerned with (1) a test's instructional relevance for a given curriculum and (2) a test's technical and practical merits.

Domain specifications connect learning outcomes to instructional content and the assessment of learning. They are developed so that someone reading them will understand instructional intentions and the means of achieving these intentions, and thus be able to develop appropriate test items. Specifications define the instructional content and the skills the teacher will teach and the student is expected to

learn. Description of instruction includes materials used, time spent on them, activities and practice for the student, and what the teacher will do. The specifications identify the content areas emphasized in instruction which will be the basis for testing at the end of instruction. Test questions are written to provide a valid sampling of learning under the conditions described. The specifications include rules for domain description, setting instructional content, generating test items, specifying test format and directions, and setting linguistic and cognitive complexity. Tests developed in this manner provide a sensitive assessment of what the student has learned and can lead to prescription responding to test diagnosis.

Item review takes place after the test has been developed. The review process deals with the extent to which a test item reflects the content of the domain; how well the item matches the domain affects the degree to which test performance is an accurate indicator of student performance. The review allows one to judge the degree to which the item belongs in the hypothetical set of items described in the specifications and how well it matches instructional content. It allows the item to be judged in terms of its fit with the domain description, content limits, item generation rules, test format and directions, and linguistic and cognitive complexity.

The procedures associated with test selection begin with the premise that a test's relevance for a given curriculum should be judged by comparing its items with specific curriculum skills in order to select the most instructionally relevant test. The procedures involve a series of judgments about one's own curriculum objectives and the degree to which these objectives are reflected in a candidate test or tests. The

procedures are especially relevant for decisions about major tests such as school- or district-wide achievement tests. The process allows test selection decisions, further, to be shared by teachers, curriculum developers, test specialists, and administrators.

Test selection should also consider a test's technical merits. The procedures in the book deal with a range of test features to be used in judging test quality; e.g., the objectives it measures, the adequacy of its development process, how it was validated, its appropriateness for the intended examinees, how it is administered, and how its scores are reported and what interpretations they permit.

Included with the book of training materials is the facilitator's guide which begins after the glossary of terms. This guide is intended for use by a workshop facilitator who will use the guide to provide training in test development and test selection. This guide (1) describes the materials in the training modules, (2) discusses the purposes of the training, and (3) deals with the advantages accruing from the domain-referenced approach to making, choosing, and using tests.

The guide provides the actual training procedures associated with each of the four modules -- (1) domain specifications, (2) item review, (3) test's relevance for a given curriculum, and (4) test's technical and practical merits.

The guide describes the step-by-step details for organizing and conducting the training. It includes the verbal instructions the facilitator provides to the participants and provides the facilitator with the necessary points to cover in discussion periods. The guide allows the training to be conducted by someone who is not necessarily an expert in tests and testing.

INTRODUCTION TO MODULE ON DOMAIN-REFERENCED TESTING

Domain-referenced test specifications define the content of a specific subject matter area and the skills or behaviors within that area which the teacher will teach and which the student is expected to learn. Test questions on the given subject are written to provide a valid sampling of student learning under the conditions described in the domain specifications. Because they are built around specific instructional content and behavioral goals, tests developed in this manner can provide a more sensitive, accurate assessment of what the learner has learned.

Domain-referenced tests require, by definition, test specifications. Test specifications insure tests that are public - both teachers and students know what will be tested and how. In other words, students do not have to spend energy guessing in order to know what to study for an upcoming test. Since test specifications form a blue print from which many items can be written, domain-referenced testing is economical - financial, psychological, and time costs are cut. In addition, since teachers can write test specifications and test items, and since the test specifications are based, right from the start, on teachers' instructional goals, domain-referenced testing is meaningful to teachers. Since test specifications counter the mystery of tests and guard against the flippant and arbitrary test maker, domain-referenced testing is also meaningful to students. Above all, domain-referenced tests are instructionally sensitive. Through solid, precise, accurate and thoughtful test specifications, domain-referenced tests can get at how much of the intended instructional content has really been learned. Domain-referenced tests, thanks to their test specifications, can help teachers answer questions about students and about teaching: Is there a way to measure

what I plan to teach? What form should that test take? How can I design a test so that I know why my students have missed a correct answer? How much of what I plan to cover will a student learn by the time of the test? Where is the best place to pick up and review material so students will master all the content by my next test.

Because a domain description is, by definition, all the possible examples and situations of behaviors, skills and knowledge in a specific content area, it is necessary to state the limits of the domain, hence Content Limits. This indicates to both the student and the teacher those items which will be stressed during the course. It will never be possible to actually test students on every situation in the domain description; each test item will be only one sample of that domain as restricted by the content limits. For these reasons, the teacher and the test-maker want to be as precise as possible in describing the domain, in writing test items congruent with the domain and of course, in providing appropriate instruction and relevant practice. The domain specification should be so clear that anyone reading it would know what instruction is implied, and therefore could use it to write test items to measure those instructional outcomes.

The following pages provide instructions for writing domain-referenced specifications, as well as examples of such specifications. The examples provided cover a variety of subjects (English, mathematics, science, social studies) and grade levels (secondary, elementary). Instructions include a description of the components of test specifications for different kinds of items.

The appendices include a list of locations where existing curriculum objectives, domain-referenced test specifications and test items may be

obtained and they also include a copy of Bloom's annotated cognitive-domain taxonomy, a classification of teaching/learning processes from simple to complex.

EXPLANATION AND SAMPLES OF DOMAIN SPECIFICATIONS

DOMAIN DESCRIPTION

A. Explanation- The domain description provides a broad, operational definition of the behavior expected of the test taker in a particular content area. This may be an objective or an explanation of a task, its components or performance conditions.

B. Samples of Domain Descriptions

1. Math -identifying shapes as triangles.
2. English-Mechanics - applying capitalization rules.
3. English-Written - writing a well organized grammatically correct paragraph in which a position is taken and supported.

CONTENT LIMITS

A. Explanation - The content limits establish the range of eligible content from which test items may be written. This may include rules for creating questions and for using prompts, cues, or additional materials (e.g., pictures, graphs, reading selections).

For Selected Response Items:
A selected response item asks the test taker to choose an answer from a number of given alternatives (e.g., true-false, matching, multiple-choice). Content limits for selected response items define and restrict the characteristics of the item stem and any additional material included in the presentation of the question or problem.

For Constructed Response Items:
Unlike the selected response item, constructed response items ask the test taker to create, not choose, and answer. Essay tests, demonstrations (e.g., driving test, cooking), drawings, oral responses, are all "constructed" responses. Content limits for constructed responses define and restrict the prompt, and where appropriate, the conditions, setting or context surrounding the testing.

B. Samples of Content Limits

1. Math - the item stem will ask the test taker to select the triangle from among four shapes, only one of which is a triangle.
2. English-Mechanics - the item stem will ask the test taker to select the word that is improperly capitalized in a given sentence. The sentence will contain at least four capitalized words, one of which is improperly capitalized according to capitalization rules.

3. Composition -- The topic presented to the students will be one with which almost all high school students would be familiar, (e.g., a topic dealing with a situation commonly encountered in daily living).

   The topic will embody an issue which permits the students to write in favor of or opposed to the proposition presented.

   One sentence will provide a brief background regarding the issue, and will explain both the pro and con positions. This sentence will be labeled: Background.

   The background sentence will be followed by the Assignment which will consist of this sentence: "Write a paragraph in which you are either in favor of, or opposed to, a (insert a brief description of one side of the issue). Be sure to support the position you have taken."

DISTRACTOR DOMAIN

A. Explanation - The distractor domain gives the wrong answers that may be used as alternatives for the selected response item. Based upon specific categories of error types, the distractor domain defines these categories of wrong answers, providing rules for generating distractors for the item.

B. Samples of Distractor Domain

   1. Math - distractors will be drawn form the set of shapes that are lacking in one of the following characteristics:
      3 sides
      straightness
      closed figures

   2. English-Mechanics - distractors will be drawn from words in the sentence that are properly capitalized according to capitalization rules.

RESPONSE CRITERIA

A. The response criteria establishes the rules and criteria for judging the quality of the test taker's generated response for the constructed response items.

B. Samples of Response Criteria

   1. Two major judgment strategies can be employed in grading the students' writing samples. The first is a separate criteria procedure, where the paragraphs are given points according to how well they meet distinct criteria (such as those to be set forth below). The second is a holistic judgment approach, where a single, overall assessment is made of each paragraph. While it is true that in the holistic approach one still employs judgmental criteria, such as how well a paragraph is organized,

these criteria are applied in a more general sense rather than in the criterion-by-criterion manner characteristic of the separate criteria approach.

2. Individuals who will be judging the paragraphs must be trained prior to their actual judging of the paragraphs. Judges should read the same paragraph, give their judgments independently, then share these judgments and discuss their reasons with the other judges. Disagreements regarding the meanings of certain criteria should be resolved. This process should be continued until judges agree on how to apply criteria to score the paragraphs.

3. During the actual judging of the paragraphs, it is desirable to have each paragraph judged independently by two judges, with a third judge being called on to resolve disagreements.

4. The following criteria might be useful in judging the paragraphs. Clearly, each would have to be explicated by the judges reading the paragraphs.

Organization

- The student has written about the assigned topic.
- The paragraph includes a topic sentence which states a position regarding the assigned topic.
- All other sentences in the paragraph support the topic sentence.

Mechanics

- Complete sentences are used rather than fragment or run-on sentences.
- Words are spelled correctly.
- Punctuation is appropriate.
- In applying the above criteria in the separate criteria approach, a predetermined number of points per criterion would be awarded to the paragraph according to how well it satisfied each criterion. For example, a 1-2-3-4-5 scale or a 1-2-3 scale might be used to indicate the extent to which the paragraph displayed acceptable spelling, punctuation, etc.

  Judges might use the above criteria in a holistic grading approach, but the criteria would be applied in an overall, rather than separate, fashion.

FORMAT

A. Explanation - The format section of the test specifications describes the lay out or form of the test.

B. Samples of Format

1. Math - multiple choice: Four shapes as response alternatives, only one of which is a proper triangle.

2. English-Mechanics - multiple choice: one sentence with four words or word groups from the sentence as response alternatives, one of which is incorrectly capitalized or left uncapitalized.

3. English-Written - constructed response: 3 paragraph expository prose prompt presented aurally and written; lined notebook paper provided for essay response.

DIRECTIONS

A. Explanation - The directions section of the test specifications provides the actual set of directions to be used or rules for generating directions.

B. Samples of Directions

1. Math - Look at the four shapes below. Only one is a triangle. Mark an X on the shape that is a triangle.

2. English-Mechanics - Select the word that is improperly capitalized.

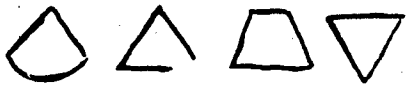
3. English-Written - In this section you must write a paragraph about an issue. In your paragraph be sure to take a pro or con position regarding the issue and support the position you have taken. Make sure your paragraph is well organized. Use complete spelling and punctuation. Write on the paper provided.

SAMPLE ITEM

A. Explanation - This section contains an example intended to guide test developers in writing items.

B. Samples of Sample Items:

1. Math - Look at the four shapes below. Only one is a triangle. Mark an X on the shape that is a triangle.



2. English-Mechanics
My Grandmother gave me a Timex watch for Christmas.
Select the word that is improperly capitalized.
   A. My
   B. Grandmother
   C. Timex
   D. Christmas

3. English-Written

Background: Some people think that there should be letter grades given for high school classes, while other people believe that all classes should be graded as either pass or fail.

Assignment: Write a paragraph in which you are either in favor of, or opposed to, a pass/fail grading system in high school.

In this section you must write a paragraph about an issue. In your paragraph be sure to take a pro or con position regarding the issue and support the position you have taken. Make sure your paragraph is well organized. Use complete sentences and correct spelling and punctuation. Write on the paper provided.

INSTRUCTIONS FOR WRITING DOMAIN SPECIFICATIONS

Before beginning to write your domain specifications there are several important decisions to make. These include decisions about the subject matter that is eligible for testing purposes, the grade level factors affecting the complexity and readibility of your items, the type of items you wish to use, and the difficulty level of the processes you wish the test takers to use in demonstrating their knowledge. Take a few minutes now to make these decisions.

- What is the general subject area or content of the test (e.g., Spanish, basic math operations, social studies skills)?

- What is the curriculum objective within this content area (e.g., conjugating common irregular Spanish verbs, solving 2-step math story problems, interpreting graphs)? You will note that the domain description, which is the first step of a domain specification, often paraphrases the curriculum objective.

- What is the grade level of the test takers (e.g., grade 6 or third-year French)? What reading level is appropriate?

- What difficulty level is desired for the test? This should be directly related to the level of instruction and type of practice test takers have been given. See a summary of Bloom's Taxonomy of the Cognitive Domain in this material (e.g., recognition or recall to facts? application or synthesis of given information?).

- What type of items do you want to use on the test? These may include selected responses, where the test taker chooses an answer from a list of alternatives (e.g., true-false, matching, multiple choice) or constructed responses, where the test taker creates the response (e.g., essay test, a demonstration, or an oral response).

Now you are ready to start writing your specifications. They should be tackled one section at a time. If you are writing a constructed response item, that is, one for which students must create their own answer rather than choose one from several alternatives, your specification will contain the following components: Domain Description, Content Limits, Response Criteria, Format, Directions, Sample Item. If your item will be a selected response item, your specification will include: Domain Description, Content LImits, Distractor Domain, Format, Directions, Sample Item.

In the following pages, space has been provided for writing each part of the domain specification. Brief descriptions of what should be included for each section head each page. For further reference, sample specifications are provided at the end of this module.

For this exercise, select one of the four objectives below to use as the domain description for your specifications.

1. Graphing a given set of data.

2. Solving mathematical word problems involving the four basic operations and numbers in decimal form.

3. Discriminating compound words from other words and dividing the compound words into their component parts.

4. Interpreting and using information on a map to answer questions.

WRITE YOUR OWN DOMAIN SPECIFICATION

Subject Area _____

Grade Level _____

Difficulty Level _____

Type of Items _____

_____

Domain Description

Definition. The domain description provides a broad, operational definition of the behavior expected of the test taker in a particular content area. This may be an objective or an explanation of a task, its components, or performance conditions.

Write your own domain description by paraphrasing the curriculum objective you have chosen.

Content Limits - Selected Responses

The content limits establish the range of eligible content from which test items may be written. This may include rules for creating questions and for using prompts, cues, or additional materials (e.g., pictures, graphs, reading selections).

A selected response item asks the test taker to choose an answer from a number of given alternatives (e.g., true-false, matching, multiple choice). Content limits for selected response items define and restrict the characteristics of the item stem and any additional material included in the presentation of the question or problem.

Content Limits - Constructed Responses

The content limits establish the range of eligible content from which test items may be written. This may include rules for creating questions and for using prompts, cues, or additional materials (e.g., pictures, graphs, reading selections).

Unlike the selected response item, constructed response items ask the test taker to create, not choose, an answer. Essay tests, demonstrations (e.g., driving test, cooking), drawings, oral responses, are all "constructed" responses. Content limits for constructed responses define and restrict the prompt, and where appropriate, the conditions, setting, or context surrounding the testing.

Distractor Domain - Selected Response Items Only

The distractor domain gives the wrong answers that may be used as alternatives for the selected response item. Based upon specific categories of error types, the distractor domain defines these categories of wrong answers, providing rules for generating distractors for the item.

Response Criteria - Constructed Response Items Only

The response criteria establish the rules and criteria for judging the quality of the test taker's generated response for the constructed response items.

Format

The format section of the test specifications describes the lay out or the form of the test.

Directions

The directions section of the test specifications provides the actual set of directions to be used or rules for generating directions.

Sample Item

The sample item is an example intended to guide test developers in writing items. Write a test item according to the rules you have outlined in the domain specification.

SAMPLE DOMAIN SPECIFICATIONS

Grade Level: Grade 5

Subject: English

Domain Description: Using correct capitalization in paragraphs adapted from a standard fifth grade test of a practical/informative nature.

Content Limits: The student will be presented with a paragraph of at least six sentences, in which all the capital letters have been omitted. Reading level should be fifth grade or lower. The test questions will consist of identifying the words which must be capitalized in a sentence from the paragraph. These words may include: the first word of a sentence; the names of languages, people, schools; days of the week; months of the year; places and buildings; titles of books or movies.

The student will be asked to correctly identify all the words in one sentence which need to be capitalized to make the sentence correct.

Distractor Domain: The distractors may include: a) omission of one word(s) within the given sentence which should be capitalized; or b) listing of a word or words in the given sentence which should not be capitalized.

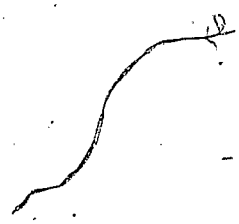
Format: Each sentence of the paragraph will be numbered. Each question will be multiple choice, with four words or groups of words listed as possible responses.

Directions: The directions will be given: "Choose the letter which lists all the capitalized words needed to make the given sentence correct."

Sample Item: 1. of all my high school friends, i remember jim the best. 2. he had a way of making adventures out of everyday events. 3. one sunday i remember in particular; it was a beautiful day in may. 4. i looked out the window, watching the sunlight dance on the columbia river. 5. my mom interrupted my daydreams, reminding me about my homework for my german class. 6. i started flipping through my history book, the american republic, to avoid beginning the german grammar. 7. suddenly a hissing voice outside the window attracted my attention. 8. it was jim; he was ready for his favorite activity, fishing. 9. we sneaked down the back stairs and out the back door.

1. In the first sentence, the following words should be capitalized:

✓ a. Of, I, Jim
b. High School
c. Of
d. Of, I



Grade Level: Grade 9

Subject: English-punctuation

Domain Description: Correctly punctuating given paragraphs adapted from a standard eighth grade text of a practical/informative nature.

Content Limits: The student will be presented with one paragraph in which all the correct punctuation marks have been omitted, except for apostrophes in contractions (I'll), possessives (Jane's), dashes and semi-colons.

For each question, students will be asked to choose all the correct punctuation marks which must be added in a given sentence to make the sentence correct. The punctuation marks to be identified and added may include:

a. periods at the end of a declarative or imperative sentence, after an abbreviation, or an initial
b. question marks following an interrogative sentence
c. exclamation point after an exclamatory sentence or interjection
d. colon after the salutation in a business letter, to separate minutes and hours in expressions of time, and before a series of things or events
e. quotation marks enclosing a quotation or a fragment of a quotation, the title of a story or poem which is part of a larger work
f. comma in a date or address; to set off words such as "yes" at the beginning of a sentence; to set off names of persons or words (phrases) in apposition; to separate words in a series, direct quotations, parallel adjectives, parenthetical phrases; after introductory prepositional phrases; before coordinate conjunctions; after the salutation and closing in a friendly letter; to separate a dependent clause from an independent clause in a complex sentence.

Distractor Domain: The distractors may include:

a. omission of necessary punctuation from the given sentence or
b. inclusion of punctuation which is not necessary or correct in the given sentence

Directions: The directions will be given: "Choose the letter of the sentence which contains all the necessary punctuation marks which will make the given sentence correct." Each sentence or group of sentences in the paragraph will be numbered.

Format: Each question will be multiple choice, with four sentence as possible responses.

Sample Item: 1. If she starts to sing again I'll crack up 2. It is funny how it hurts to hold back a laugh 3. I was sitting in the auditorium at 10:00 am and we were having a singing rehearsal for graduation 4. Sit up Get off those shoulders Think tall Sing tall Sing like this said Ms Small 5. I knew that if she was going to tweet like a bird again I would laugh 6. But I just could not laugh because Ms Small would kick me out of the auditorium and that meant Felson's office--and no graduation 7. La la la--sing children Sing with your hearts said Ms Small 8. I couldn't hold it 9. She was so funny I almost rolled of the auditorium seat 10. The other students didn't laugh, but me I sounded like Santa Claus 11. It became quiet for a second 12. What are you doing Joe I know it is you Present yourself to Mr Felson at once that voice said 13. Ms Small is a foot shorter than a tall Coke but she has the bark of a hungry hound dog

1. The first sentence should be written:

a. If she starts to sing again I'll crack up.
b. If she, starts to sing again, I'll crack up
✓ c. If she starts to sing again, I'll crack up.
d. If she starts, to sing again, I'll crack up.

Grade Level: Grade 8

Subject: Introduction to Algebra

Domain Description: Using four basic arithmetic operations and the properties of equations and inequalities determine solution set of linear open sentences with one unknown quantity.

Content Limits:

1. Stimuli include a number sentence with one unknown quantity, represented by a lower case letter in italics, and an array of five solution sets or single answers, only one of which is correct.

2. Number sentences may be statements of equalities or inequalities.

3. The number sentence may require the use of any of the following properties in its solution: adding or subtracting equal quantities from both sides, multiplying or dividing both sides by equal positive quantities, multiplying or dividing both sides by equal negative quantities.

4. Factoring may be a requisite operation for solving the equation.

5. Application of the distributive property of multiplication may be required for solving the equation.

6. Number sentences will have no more than five terms. Both fractions and decimals may be used, but not in the same expression. Terms with exponents (powers) may appear in the number sentence only if they cancel out and need not be expanded. No higher powers may be used.

7. Solution sets for equations and inequalities will be drawn from the set of positive and negative rational numbers. The null set (Ø) may also be used as a correct solution set.

8. The solution set for a particular number sentence may be drawn from the set of integers, or the set of positive integers, if it is stated that the unknown quantity in that particular number sentence is an integer or a positive integer.

Distractor Domain:

1. Distractors may be drawn from the set of wrong answers resulting from errors involving any of the properties discussed in 3, 4, or 5 above.

2. Distractors may also be drawn from the set of wrong answers due to incomplete solution sets.

3. Distractors may not reflect errors due to wild guessing.

4. "None of the above" is not an acceptable alternative.

Format: The equation with one unknown will be presented. Five response alternatives, the correct response and four distractors, will be listed below the equation.

Directions: Solve the equation. Then select the correct answer or solution set from the choices given.

Sample Item:

1. $8n + 2 = 2n + 38$; $n = ?$

a) $n = 3$
✓ b) $n = 6$
c) $n = 4$
d) $n = 5$
e) $n = 7.6$

2. If x is an integer and $16x \leq 32$; $x = ?$

a) $x = 48$
✓ b) $x \in \{..0, 1, 2\}$
c) $x = 2$
d) $x \leq \emptyset$
e) $x \in \{3, 4, 5...\}$

Grade Level: Secondary

Subject: Life science - circulatory system

Domain Description: Recognizing and differentiating the structures and functions of each of the circulatory systems.

Content Limits:

1. Circulatory systems include: pulmonary circulation, coronary circulation, systemic circulation (renal and portal).

   Heart structures eligible for identification and differentiation of function include: left and right atria (or auricles), left and right ventricles, pulmonary artery and veins, systemic artery and veins, aorta, valves.

   Other structures eligible: veins, arteries, capillaries, femoral artery and vein, inferior vena cava and superior vena cava, jugular vein and carotid artery, brachial artery, and basilic vein, portal and renal veins and arteries.

2. In items requiring labelling, a list of terms should be provided including all correct terms and additional relevant terms, from which the test taker may select labels to use.

Distractor Domain:

1. Distractors should represent misidentification of terms, functions.

2. Distractors may include responses that are incorrect because they are incomplete or inadequate.

Format: Each question will be multiple choice, with four response alternatives, three distractors and one correct response.

Directions: Select the one correct answer.

Sample Item: Select the one correct answer.

1. _____ assist the heart in pumping blood by constricting and expanding as blood is pumped into them.

   a) Veins
   b) Capillaries
✓ c) Arteries
   d) Valves

Grade Level: Secondary

Subject: Life science - circulatory system

Domain Description: Applying understanding of the circulation system to predict cause-effect relationships within the system.

Content Limits:

1. Circulatory systems include: pulmonary circulation, coronary circulation, systemic circulation (renal and portal).

   Heart structures eligible for identification and differentiation of function include: left and right atria (or auricles), left and right ventricles, pulmonary artery and veins, systemic artery and veins, aorta, valves.

   Other structures eligible: veins, arteries, capillaries, femoral artery and vein, inferior vena cava and superior vena cava, jugular vein and carotid artery, brachial artery and basilic vein, portal and renal veins and arteries.

   Eligible cause-effect situations include: heart attack, arteriosclerosis, injury to aorta or other major veins and arteries (superior, inferior vena cava, jugular, carotid, femoral veins/arteries, portal and renal veins and arteries, brachial and basilic), high blood pressure, pulse, heart murmur.

2. Items on cause-effect may present the cause and ask the effect or vice versa. These items may be presented pictorially, (e.g., showing a blood clot in the coronary artery). However, in these cases, all parts must be labelled for the student.

Distractor Domain:

1. Distractors should represent misidentification of terms, functions.

2. Distractors may include responses that are incorrect because they are incomplete or inadequate for items concerned with processes and systems only.

OR

Response Criteria:

1. For labelling pictures, terms must be correct; spelling does not count. Partial credit may be given for correct

labels in pictures requiring more than one response; incorrect labelling that affects meaning (e.g., not including the word artery or vein as in carotid), should be counted as incorrect.

2. Correct responses to the cause-effect constructed responses must include all underlined points below. Partial credit may be awarded at the discretion of the teacher.

   a. heart attack: clot in coronary artery preventing the flow of blood to the heart; heart tissue damaged or destroyed ude to lack of food and oxygen since blood can't reach cells.

   b. injury to major veins and arteries: should differentiate the functions and locations of the given vein or artery (femoral artery and vein; inferior and superior vena cava; jugular vein and carotid artery; brachial artery and basilic vein; portal and renal veins and arteries; aorta).

   c. arteriosclerosis: described as loss of elasticity of artery walls which normally stretch and relax with the pulsing during heartbeat. Lost elasticity, often due to fatty deposits on the artery walls (hardening of the arteries), can create abnormally high blood pressure as the blood is pushed through narrower ducts.

   d. high blood pressure: could describe two possible causes--exercise (heart pumps harder to supply more oxygen to the muscles), and changes to the blood vessels (e.g., arteriosclerosis - smaller tube way for blood flow increases pressure).

   e. pulse and heartbeat: should describe the pumping action of the heart as reflected in the arteries, stretching the arterial walls, pulse as accurate indicator of heart action.

   f. heart murmur: must describe valve functions, normally and their sound (ventricles contract and valves close; ventricles relax and aorta valves close). Murmur represents backflow of blood from incomplete or improper valve closing.

Format:
Selected Response:

The question will be multiple choice with four alternatives, three distractors and one correct response.

OR

Constructed Response:	The question will require filling in blanks, labelling figures, or writing a paragraph.
Directions: Selected Reponse Items:	For multiple choice items -- select the one correct answer.
	OR
Constructed Response Items:	Complete each sentence. OR Label each part of the diagram representing_____. OR Diagram (or describe) the _____ process through the heart. OR Answer each question completely, including a description of causes, effects, parts, functions or processes where necessary.
Constructed Response Sample Item:	Answer completely, including a description of parts or functions where necessary. What would be the effect of injury to the carotid artery?

Grade Level: Secondary

Subject: Life Science - circulatory system

Domain Description: Explaining/describing the process of circulation (pictorial and verbal).

Content Limits:

1. Circulatory systems include: pulmonary circulation, coronary circulation, systemic circulation (renal and portal).

   Heart structures eligible for identification and differentiation of function include: left and right atria (or auricles), left and right ventricles, pulmonary artery and veins, systemic artery and veins, aorta, valves.

   Other structures eligible: veins, arteries, capillaries, femoral artery and vein, inferior vena cava and superior vena cava, jugular vein and carotid artery, brachial artery and basilic vein, portal and renal veins and arteries.

   Eligible cause-effect situations include: heart attack, arteriosclerosis, injury to aorta or other major veins and arteries (superior, inferior vena cava, jugular, carotid, femoral veins/arteries, portal and renal veins and arteries, brachial and basilic), high blood pressure, pulse, heart murmur.

2. In items requiring diagramming, basic representations should be provided so that test takers need only supply labels and arrows.

3. In items requiring labelling, a list of terms should be provided including all correct terms and additional relevant terms, from which the test taker may select labels to use (unless recall is being tested).

Distractor Domain:

1. Distractors should represent misidentification of terms, functions.

2. Distractors may include responses that are incorrect because they are incomplete or inadequate for items concerned with processes and systems.

OR

Response Criteria:

1. For labelling pictures, terms must be correct; spelling does not count. Partial credit may be given for correct

labels in pictures requiring more than one response; incorrect labelling that affects meaning (e.g.,,not including the word artery or vein as in carotid), should be counted as incorrect.

2. Correct responses to the cause-effect constructed responses must include all underlined points below. Partial credit may be awarded at the discretion of the teacher.

   a. heart attack: clot in coronary artery preventing the flow of blood to the heart; heart tissue damaged or destroyed due to lack of food and oxygen since blood cannot reach cells.

   b. injury to major veins and arteries: should differentiate the functions and locations of the given vein or artery (femoral artery and vein; inferior and superior vena cava; jugular vein and carotid artery; brachial artery and basilic vein; portal and renal veins and arteries; aorta).

   c. arteriosclerosis: described as loss of elasticity of artery walls which normally stretch and relax with the pulsing during heartbeat. Lost elasticity, often due to fatty deposits on the artery walls (hardening of the arteries), can create abnormally high blood pressure as the blood is pushed through narrower ducts.

   d. high blood pressure: could describe two possible causes--exercise (heart pumps harder to supply more oxygen to the muscles), and changes to the blood vessels (e.g., arteriosclerosis-smaller tube way for blood flow increases pressure).

   e. pulse and heartbeat: should describe the pumping action of the heart as reflected in the arteries, stretching the arterial walls, pulse as accurate indicator of heart action.

   f. heart murmur: must describe valve functions, normally and their sound (ventricles contract and valves close; ventricles relax and aorta valves close). Murmur represents backflow of blood from incomplete or improper valve closing.

Format:
Selected Response: The question will be multiple choice with four response alternatives, three distractors and one correct response.

OR

Constructed Response: The question will require filling in blanks, labeling figures or writing a paragraph.

Directions: Selected Response: Select the one correct answer.

OR

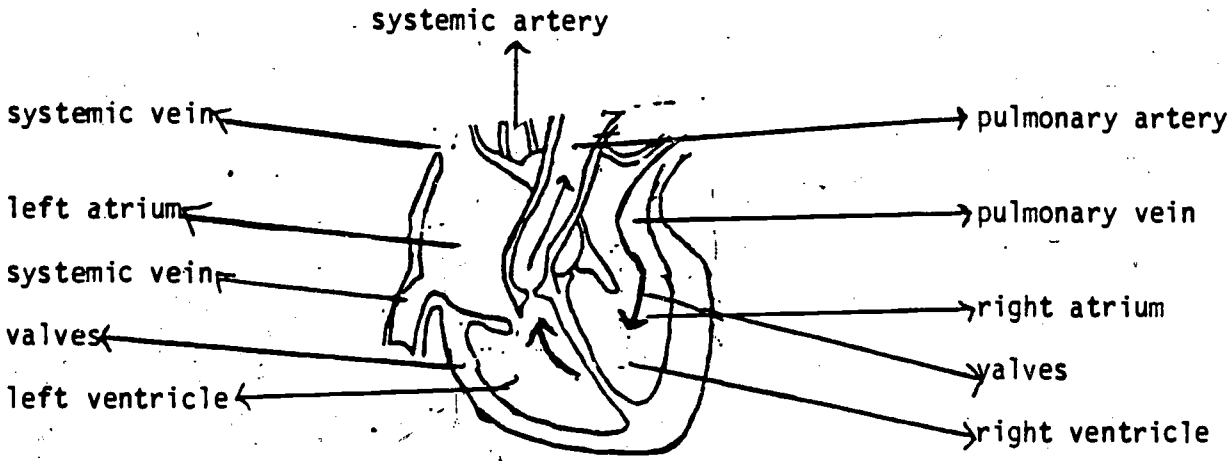
Constructed Response: Complete each sentence. OR Label each part of the diagram representing _____. Or Diagram (or describe) the _____ process through the heart. OR Answer each question completely, including a description of causes, effects, functions, parts, and processes as necessary.

Selected Response Sample Item: Select one correct answer.

1. _____ assist the heart in pumping blood by constricting and expanding as blood is pumped into them.

   a. Veins
   b. Capillaries
   ✓ c. Arteries
   d. Valves

Constructed Response Sample Item: Label each part of the diagram representing the path of the blood through the pulmonary cirulatory system. Draw arrows and use labels from the list provided.



pulmonary vein
pulmonary artery
aorta
left auricle/atrium
right auricle/atrium
valve(s)
systemic vein(s)
systemic artery
left ventricle
right ventricle
renal artery
portal vein
renal vein
portal artery
capillary

AN ANNOTATED COGNITIVE DOMAIN TAXONOMY*

This classification describes, from simplest to most complex, six degrees to which information that is taught can be learned.

1. Knowledge. Recalling information pretty much as it was learned.
   In its simplest manifestation, this includes terms, facts, dates and names - associated with a subject matter area. At a more complex level, it means knowing the major sub-areas, methods of inquiry, classifications and ways of thinking characteristic of the subject area, as well as its central theories and principles.

2. Comprehension. Reporting information in a way other than how it was learned in order to show that it has been understood.
   Most basically this means reporting something learned through an alternative medium. More complex evidence of comprehension involves interpreting information in "one's own words" or in some other original way, or extrapolating from it to new but related ideas and implications.

3. Application. Use of learned information to solve a problem.
   This means carrying over knowledge of facts or methods learned in one specific context to a new context.

4. Analysis. Taking learned information apart.
   Analysis means figuring out a subject matter's most elemental ideas and their interrelationships.

5. Synthesis. Creating something new and good, based on some criterion.
   This creation can be something that communicates to an audience, that plans a successful goal-directed endeavor, or that subsumes a collection of ideas within a new theory.

6. Evaluation. Judging the value of something for a particular purpose.
   This means making a statement of something's worth based either on one's own well-developed criteria or on the well-understood criteria of another.

---

* Adapted from TAXONOMY OF EDUCATIONAL OBJECTIVES: The Classification of Educational Goals: HANDBOOK 1: Cognitive Domain, by Benjamin S. Bloom, et al. Copyright 1956 by Longman Inc. Previously published by David McKay Company, Inc. By permission of Longman Inc.

Appendix B
Module 1

Sources of Measurable Objectives in a Variety of Subject Areas

Source	Address
Clark County Curriculum Guides	Clark County School District 2832 East Flamingo Road Las Vegas, Nevada 89121
Course Goals Developed by the Tri-County Project	Commercial-Educational Distributing Services P.O. Box 8723 Portland, Oregon 97208
Elective Quarter Plan Curriculum Materials, Grades 1-12	Director of Curriculum Jefferson County Board of Education 3023 Melbourne Avenue Louisville, Kentucky 40220
Evaluation for Individualized Instruction (EII) Pool of Behavioral Objectives and Test Items for K-12 in Language Arts, Math, Social Science, and Science	Institute for Educational Research 1400 West Maple Avenue Downers Grove, Illinois 60515
Individual Pupil Monitoring System (IPMS) Behavioral Objectives Booklets for Grades 1-6 in Reading and Grades 1-8 in Mathematics	The Test Department Houghton-Mifflin Company 777 California Avenue Palo Alto, California 94304
Learning Objectives and Behavioral Objectives (Primary, Secondary, Junior College)	Cambridge Book Company 488 Madison Avenue New York, New York 10022

Appendix B
Module 1 continued

Source	Address
Measurable Objectives Collection in many Subjects for Grades K-12	Instructional Objectives Exchange Box 24095 Los Angeles, California 90024
Objectives and Items for K-12 in Language Arts, Math, Social Sciences, Science, and Vocational Education	The Co-op 413 Hills House North University of Massachusetts Amherst, Massachusetts 01002
Specimen Set of Mastery; an Evaluation Tool for Reading and Math, Grades K-9	Science Research Associates 259 East Erie Street Chicago, Illinois 60611

Appendix B
Module 1 continued

Sources of Broadly-Stated Goals and Goal Categories

Source	Address
Brochures of Objective in Art, Career and Occupational Development, Literature, Mathematics, Music Reading, Science, Social Studies and Writing	National Assessment of Educational Progress 600 Lincoln Tower 1860 Lincoln Street Denver, Colorado 80203
Taxonomy of Educational Objectives, Handbooks I & II	David McKay Co., Inc. 750 Third Avenue New York, New York 10017
Workshop Packet for Educational Goals and Objectives	Phi Delta Kappa, Inc. Eighth Street & Union Avenue Box 789 Bloomington, Indiana 47401

INTRODUCTION TO THE ITEM RATING SCALE - MODULE II

Background

Domain-referenced testing is based on the assumption that by limiting and defining a class of behaviors, skills and information (a domain), a set of rules (test specifications) may be created for generating actual test items. The degree to which these items reflect the content of the domain affects the degree to which test performance accurately indicates competence in the domain. This test feature, descriptive validity, is an important consideration in selecting or creating tests. Afterall, the most thoughtfully written test specifications are only helpful in pin pointing student competence if they are translated accurately into test items. The Item Rating Scale (IRS) was developed to provide a systematic and reliable method of judging the descriptive validity of test items and their specifications.

The IRS methodology provides a continuum of values to use in judging the "belongingness" of an element to a set. This kind of judgment involves examing the rules governing membership in the set to determine how well a specific element represents the whole set. In otherwords, these judgments are probability statements describing the likelihood that the given element is a member of the given set.

In applying this concept to judgments of descriptive validity for domain-referenced tests, the "set" is the hypothetical set of items described in the test specifications, these specifications are the "rules" governing membership in the set, and the "element" is the test item. Judgments are made about how well an item reflects the assessment intentions of its test specification; that is, how well the item measures the domain. This

judgment is not a yes/no choice. That is, each of the several features, or dimensions, of test items that are elaborated in a test specification (e.g., distractor limits, format) may affect how well an item matches the test specifications for a domain. The IRS has been devised to permit judgments to be made about item compatability with each of the categories in test specifications.

Description

The IRS, which presumes the availability of test specifications, is used in an item-by-item review of a test or a group of test items. The specifications may be those that accompany a particular test or those that have been locally developed. In either case, once a set of test specifications has been developed, the next task is to assemble test items that match these guidelines or intentions. This careful matching increases the likelyhood that the test will provide a valid assessment of student performance in the content area under the conditions described in the specifications.

The Item Review Scale, then, helps in judging the probability that any given item is a legitimate member of the hypothetical set of items defined by the test specifications. More specifically, the IRS is used to judge, along eight independent categories or dimensions the probability of match between the test specifications and any given item.

The first six rating categories of the IRS parallel the structure of domain-referenced test specifications. These categories consist of: Domain Description, Content Limits, Distractor Limits or Response Criteria, Format, Directions, and Sample Item. In addition, two other categories, linguistic and thinking complexity, are also included in the IRS since they affect the

match between the item and the test intentions embodied in the specifications.

The first category of the IRS concerns the general Domain Description. The second category, Content Limits, compares the description of eligible subject matter and item features with the test item's content and features. The third category is either the Distractor Limits or Response Criteria, depending on whether the item is a selected or constructed response type. For selected response items, the specification rules for creating wrong answer alternatives are compared with the actual wrong answer choices used in the test item. For constructed response items, the prescribed criteria for evaluating the examinee's response are compared both to those criteria used and to the suitability of the item and the test conditions. Format and Directions are the fourth and fifth categories to match specifications and actual items. Here, the concern is whether the layout of the item and the directions for completing the test conform to the test specifications. The Sample Item is the final aspect of the test specifications included in the Item Rating Scale.

The two additional categories not necessarily included in the test specifications, Linguistic Complexity and Thinking Compelxity, provide more information about how student performance might differ. These biasing elements are important to the degree that the specifications and resulting items are intended to provide the same measure of performance for all students in the given area.

How to Use the IRS

Each category described above appears in the IRS in boldface, followed by several statements describing test item features that are indicators of

item match with the test specification component under consideration. Raters are asked to use these statements to judge test items against their specifications.

Raters then assign a whole number value, from 0 to 10 inclusive, that best represents their judgment of the probability that the item and specification belong together on each particular dimension (e.g., Domain Description). In this 0 to 10 scale, 0 indicates a highly improbable match between item and specification and 10 indicates a highly probable match. Raters proceed one category or dimension at a time. To assist raters in their judgment, sample items are presented opposite each IRS category illustrating high and low probability ratings.

When all eight categories have been individually scored, overall probability rating may be calculated for the item. The final calculations are guided by the Overall Item Rating Scale which applies a weighting system to incorporate the scores in each category.

Interpretations of the ratings are offered in terms of the three features judged to be most critical -- content limits, distractor domain or response criteria, and thinking complexity. Implications for item revision or, where necessary, specification revision, are also briefly stated. This information will allow for more reliable, confident decisions by test makers to use, modify, or reject particular test items.

DIRECTIONS

The Item Rating Scale (IRS) is intended for use in making systematic content validity judgments for domain-referenced tests by comparing test specifications with test items. The Scale also provides feedback for revising items or specifications as necessary. In using the IRS, one test item at a time is rated against a set of test specifications.

1. Get a copy of the test specifications and the items you wish to rate.

2. Go through the categories of the IRS using the statements in each section to direct you in judging the compatability of your item with the six test specification features and the two additional categories concerned with complexity issues.

3. In each section, rate the probability that your item is a member of the hypothetical set of items described by the test specifications in that category. Use a scale of 0 to 10 to rate your item, letting 0 indicate a highly improbable match and 10 a highly probable one.

   The following guidelines are suggested for assigning number ratings in each section:

   0,1,2 *This rating range should be used for items that are completely unrelated to the specification in the dimension you are rating.*

   3,4,5 *This rating range should be used for items that are vaguely related and/or inadequate.*

   6,7 *This rating range should be used for items you feel would definitely require a second look and some revision, but which you feel reluctant to totally abandon.*

   8,9 *This rating range should be used for items that you feel are good representative match-ups with the specifications although slightly off.*

   10 *This rating should be used for items that are beyond a doubt perfect examples of the specification.*

   Enter your rating in the box provided.

   Space for taking notes has been provided with each section or category. It is strongly suggested that you take advantage of this to make comments about the item as you rate it. Such notes will be useful later in revising the item or the specifications.

4. Complete the Overall Item Rating sheet by carrying over the rating scores from each section to the appropriate line of the rating sheet. Make the calculations indicated in the directions there, applying the rating weights where indicated.

5. Refer to the Interpretation Guide for rating explanations.

6. REMEMBER YOU ARE RATING THE MATCH BETWEEN THE ITEM AND THE SPECIFICATION, NOT THE ITEM AND YOUR EXPECTATIONS OR STANDARDS! ALSO, EACH IRS CATEGORY SHOULD BE RATED INDEPENDENTLY OF THE OTHERS, FOR EXAMPLE, DOMAIN DESCRIPTION RATINGS DO NOT INCLUDE CONTENT LIMIT CONSIDERATIONS. USE THE STATEMENTS PROVIDED TO GUIDE YOUR JUDGMENTS.

Item Rating Scale
Description of Categories

I. DOMAIN DESCRIPTION

1. The test item is a good and fair representative of the subject area outlined in the domain description of the test specifications. It does not assess an obscure or unusual aspect of the domain.

2. Test item conditions are not at odds with test intentions. This is especially important in constructed items.

3. The test item content is closely related to the instructional objective(s) stated or implied in the domain description.

II. CONTENT LIMITS--SELECTED RESPONSE ITEMS ONLY

1. The item and additional accompanying material (e.g., graphs, maps, reading selections) follow the content limits on length and general difficulty level.

2. The item and additional accompanying material follow the content limits on eligible content, descriptive detail and completeness of information provided.

3. The solution processes required to answer the item match those described or implied in the content limits.

II. CONTENT LIMITS--CONSTRUCTED RESPONSE ITEMS ONLY

1. The item matches the content limits on eligible content, descriptive detail, or completeness of the prompting information provided.

2. The item provides a context for responding that is similar to that described in the content limits (e.g., time restrictions, length of written/oral response, equipment or aid restrictions, warmup or false start provisions).

3. The mental processes required to respond to the item seem to match those described or implied in the content limits.

III. DISTRACTOR LIMITS--SELECTED RESPONSE ITEMS ONLY

1. The alternative answers, or distractors, provided in the item require the test taker to discriminate important features or factors described in the distractor domain as differentiating correct from incorrect answers. Distinctions between correct and incorrect answers are not based on trivial or irrelevant features.

2. The distractors provided in the item correspond to the content limits on number, length, and general level of difficulty.

III. RESPONSE CRITERIA--CONSTRUCTED RESPONSE ITEMS ONLY

1. The rules used to judge the student's response are those described by the response criteria.

2. The item prompt sets up a context for responding that is appropriate to the response criteria for judging the content and style/form of the response (i.e., likely to elicit a judgeable response).

3. Problems arising from incomplete or inadequate answers are dealt with in a way that upholds the testing intentions of the specifications.

IV. FORMAT

1. The organization and display (layout) of the item conforms to the format description in the test specifications.

2. FOR SELECTED RESPONSE ITEMS ONLY: The organization and display of any additional information (e.g., maps, graphs, pictures, reading selections) conforms to the format description.

   FOR CONSTRUCTED RESPONSE ITEMS ONLY: The context or conditions for responding to the item (e.g., time limits, space limits, available equipment) conform to the format description.

V. DIRECTIONS

1. The directions for completing the test item correspond to the description of test directions in the test specifications.

2. The reading level and complexity of the directions follow the description of test directions in the test specifications; or seem to be within suitable range for the intended test takers.

VI. SAMPLE ITEM

1. The sample item and the test item being rated could come from the same set of items described by the test specifications.

2. The sample item and the test item are very similar in content and either distractors or response criteria.

3. The sample item and the test item are very similar in format and directions.

LINGUISTIC COMPLEXITY

1. Vocabulary used in the item is consistent with the test specifications for item difficulty. Words are not used that have different or unfamiliar meanings for different students or student groups.

2. Item language structure (e.g., the use of compound, complex sentences, antecedents) is consistent with the test specifications for item difficulty.

VIII. THINKING COMPLEXITY

1. Those mental processes required for the solution or performance of the test item, but that are not described in the domain description or content limits (i.e., are assumed) are readily available to all test takers at some necessary level of competence (e.g., drawing ability, handwriting legibility, short-term memory capacity, imagination, ability to separate relevant from irrelevant, detail from generalization).

2. Directions for completing the test item provide the same amount of information and structure for all test takers. Everyone has the same understanding of what is expected and of what the limits or rules for answering are.

3. FOR ITEMS WITH NONVERBAL COMPONENTS, it is reasonable to assume that these components conform with the content limits or distractor domain in their intended meaning, and that this interpretation is stable across all groups of test takers.

OVERALL ITEM RATING

1. Recopy item ratings from each section, making the indicated weighting adjustments for the starred features: Content Limits, Distractor Limits or Response Criteria, and Thinking Complexity.

DOMAIN DESCRIPTION		_____
*CONTENT LIMITS	(_____ x 3)	=_____
*DISTRACTOR LIMITS OR RESPONSE CRITERIA	(_____ x 3)	=_____
FORMAT		_____
DIRECTIONS		_____
SAMPLE ITEM		_____
LINGUISTIC COMPLEXITY		_____
*THINKING COMPLEXITY	(_____ x 3)	=_____
	TOTAL	_____

2. Total the scores. Divide the total by 14. This number is the overall item rating.

OVERALL ITEM RATING _____ ÷ 14 =_____

3. Refer to the Interpretation Guide for assistance in making decisions about the item and for suggestions for modifying the item according to its rating.

IRS INTERPRETATION GUIDE

ITEMS RATED 7 OR BETTER

IF ALL THREE STARRED CRITICAL FEATURES ARE RATED 8 OR BETTER*, your item is good, basically in conformity with the test specifications. Review and rewrite efforts should be directed toward other features that scored low, (e.g., Format). Use the statements in the IRS rating categories to guide your work.

IF ONE CRITICAL FEATURE RECEIVED A RATING OF 7 OR LOWER*, go back to the specifications on that feature. Try to better align your item with the testing intentions described in the specifications. Use the statements in the IRS to help direct your thinking. You also have problems with other features. Rewrite the item but review it again to be certain all critical features are up to par.

IF MORE THAN ONE CRITICAL FEATURE RECEIVED A RATING OF 7 OR LOWER*, the item has serious validity problems. If this is the kind of test item you want, then you should reconsider the specifications you are using. They may need to be better conceptualized, reconceptualized, or more complete in their description of item qualities. If the specifications are closer to what you want to be testing, throw out the item. Find or write a new item.

ITEMS RATED BELOW 7

IF ALL THREE STARRED CRITICAL FEATURES ARE RATED 8 OR BETTER*, your item is potentially a good item but has serious problems in presentation. Go back to the specifications for those features receiving the low ratings. Clean up your item. Use the statements in the IRS rating categories to guide your efforts.

IF ONE OR MORE OF THE CRITICAL FEATURES SCORED 7 OR LOWER*, your item isn't worth the fix-up effort. Before you start over, reconsider the specifications with which you are working; they may need to be better conceptualized or more complete in their description of item features.

* Before rating weights are applied.

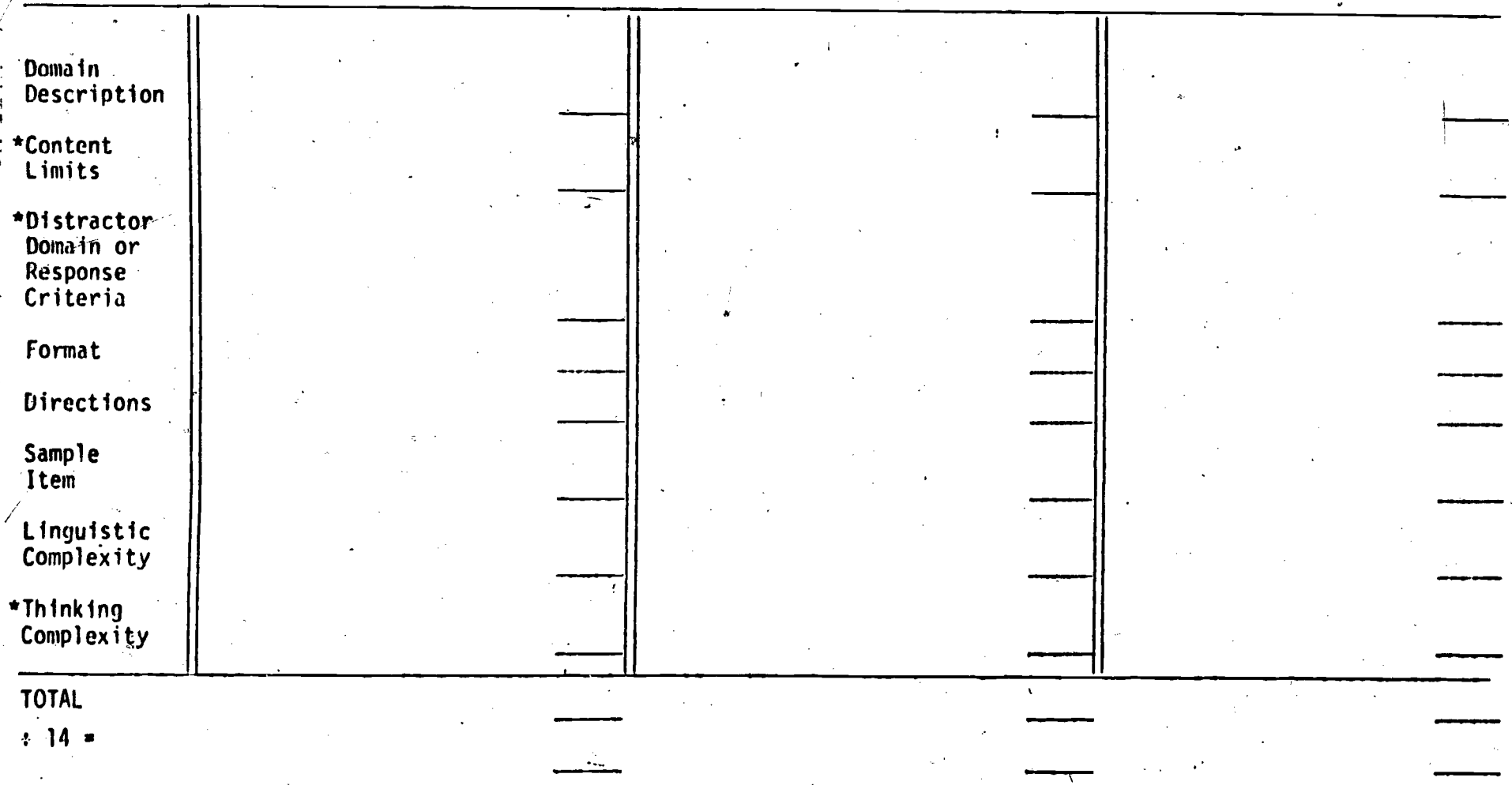
Item Rating Form

SPECIFICATION BEING RATED _____

RATER TITLE _____

COMMENTS: (additional comments can be made on the reverse side)

RATING SCALE



*Critical features

2.12

ENGLISH-PUNCTUATION

Grade Level: Grade 7 and 8

Subject: English-Punctuation

Domain Description: Applying the rules of punctuation to correctly punctuate given prose material missing end punctuation, commas and quotation marks.

Content Limits:

1. Students will be presented with a passage containing six to twelve sentences at a sixth grade level. The following punctuation marks will be omitted:

   a. periods at the end of statements
   b. question marks at the end of interrogative sentences
   c. exclamation marks after exclamatory sentences or interjections or commands
   d. quotation marks enclosing a speaker's or character's words
   e. commas to set off names of persons, or items in a series, to set off words such as "yes," "no," "well," "however," "meanwhile;" to separate parenthetical phrases, to precede coordinate conjunctions in a long or compound sentences, and at the beginning or end of a speaker's quote, such as "Come in," said Ben.

2. Any other punctuation marks already in the selection and all other parts of the selection (grammar, capitalization, etc.).

3. Each sentence of the selection will be numbered and questions on a given sentence will refer to those numbers. All or some of the sentences of the passage may be used as questions.

Distractor Domain: The distractors will include 1) ommission of punctuation marks or 2) inclusion of punctuation marks which are not necessary or are incorrect. Only the punctuation marks listed above will be used in the distractor domain (e.g., no semicolons or colons).

Format: A prose passage will be given. The sentences will be numbered. Multiple choice questions will consist of a stem and four alternative responses, three distractors and the correct response.

Directions: Students will be asked to choose the answer that correctly punctuates the sentence.

Sample Item:

1. As Tom was whitewashing the fence his friend Ben walked by 2. Ben stared a moment 3. Then he said Hi Can you go fishing 3. There was no answer 4. Tom stepped back to note the effect and added a touch here and there 5. Meanwhile Ben was watching every move and getting more and more interested 6. Say Tom let me whitewash a little Ben said 7. Tom was about to agree but he changed his mind 8. No said Tom it's got to be done very carefully 9. I reckon there ain't but one boy in a thousand maybe two thousand that can do it the way it's got to be done

1. The first sentence should read:

   a. As Tom was whitewashing the fence his friend, Ben walked by.
   b. As Tom was whitewashing the fence, his friend Ben, walked by.
   ✓c. As Tom was whitewashing the fence, his friend, Ben, walked by.
   d. As Tom was whitewashing the fence his friend Ben, walked by.

USE THE FOLLOWING PARAGRAPH TO ANSWER THE QUESTIONS BELOW:

1. The loudspeakers boomed Four three two and a deep rumble began to come from the tail of the rocket 2. All the men in the control tower looked at their instruments 3. All systems were still Go 4. One zero lift-off said the voice on the speaker system 5. A huge roar shook the ground as the rocket Enterprise II began to move slowly off the ground 6. Some of the men cheered Go baby go

Choose the answer which contains all the necessary and appropriate punctuation marks. Write the letter of the correct answer on your answer sheet.

1. Sentence number 1 should be written
   ✓ a) The loudspeakers boomed, "Four, three, two," and a deep rumble began to come from the tail of the rocket.
   b) The loudspeakers boomed "Four, three, two" and a deep rumble began to come from the tail of the rocket.
   c) The loudspeakers boomed. Four, three, two and a deep rumble began to come from the tail of the rocket.
   d) No punctuation is necessary.

2. Sentence number 5 should be written
   a) A huge roar shook the ground, as the rocket Enterprise II began to move slowly off the ground.
   ✓ b) A huge roar shook the gound as the rocket, Enterprise II, began to move slowly off the ground.
   c) A huge roar shook the ground as the rocket "Enterprise II" began to move slowly off the ground.
   d) A huge roar shook the ground as the rocket, Enterprise II began to move slowly off the ground.

3. Sentence number 6 should be written
   a) Some of the men cheered "Go baby go!"
   b) Some of the men cheered, "Go baby go!"
   ✓ c) Some of the men cheered, "Go, baby, go!"
   d) Some of the men cheered, "Go, baby, go".

USE THE FOLLOWING PARAGRAPH TO ANSWER THE QUESTIONS BELOW:

1. Milton the city banker claimed he saw a flying saucer near the lake last night 2. Looked like a giant two story bell reported Milton 3. The silver saucer was round near the top and triangular shaped near the bottom 4. Near the top were three round windows 5. I think I saw some space creatures looking through these windows 6. Close to the bottom were four large square windows and a seven foot door 7. When giant flames shot out of the bottom of the saucer the ship moved up and down

Circle the letter of the answer which contains all the necessary punctuation marks.

4. Sentence number 1 should be written
   a) Milton, the city banker, claimed he saw a flying saucer near the Lake last night.
   b) Milton, the city banker, claimed he saw a flying saucer near the lake, last night.
   c) Milton, the city banker claimed he saw a flying saucer near the lake last night.
   d) Milton, the city banker, claimed he saw a flying saucer near the lake last night.

5. Sentence number 4 should be written
   a) Near the top, were three round windows.
   b) Near the top were three, round windows.
   c) Near the top were three round windows.
   d) "Near the top were three round windows."
   e) "Near the top, were three round windows".

6. Sentence number 7 should be read
   a) When giant flames shot out of the bottom of the saucer the ship moved up and down.
   b) When giant flames shot out of the bottom, of the saucer, the ship moved up and down.
   c) When giant flames shot out of the bottom of the Saucer, the ship moved up and down!
   d) When giant flames shot out of the bottom of the saucer, the ship moved up and down.

USE THE FOLLOWING PARAGRAPH TO ANSWER THE QUESTIONS BELOW:

1. what is a koala bear 2. it is a small animal that lives in the trees of Australia. 3. a koala bear looks like a teddy bear 4. He has a big head and a short nose. 5. a koala bear is about two feet long one foot high and has a little tail

Circle the letter of the best answer.

7. Sentence number 1 should be written
   a) What is a koala bear.
   b) What is a Koala Bear.
   c) what is a koala bear?
   d) What is a koala bear?

8. What is wrong with sentence number 4?
   a) It should have an exclamation mark (!).
   b) It should have a comma after the word "head."
   c) It should have a comma after the word "big."
   d) No punctuation is necessary; the sentence is okay.

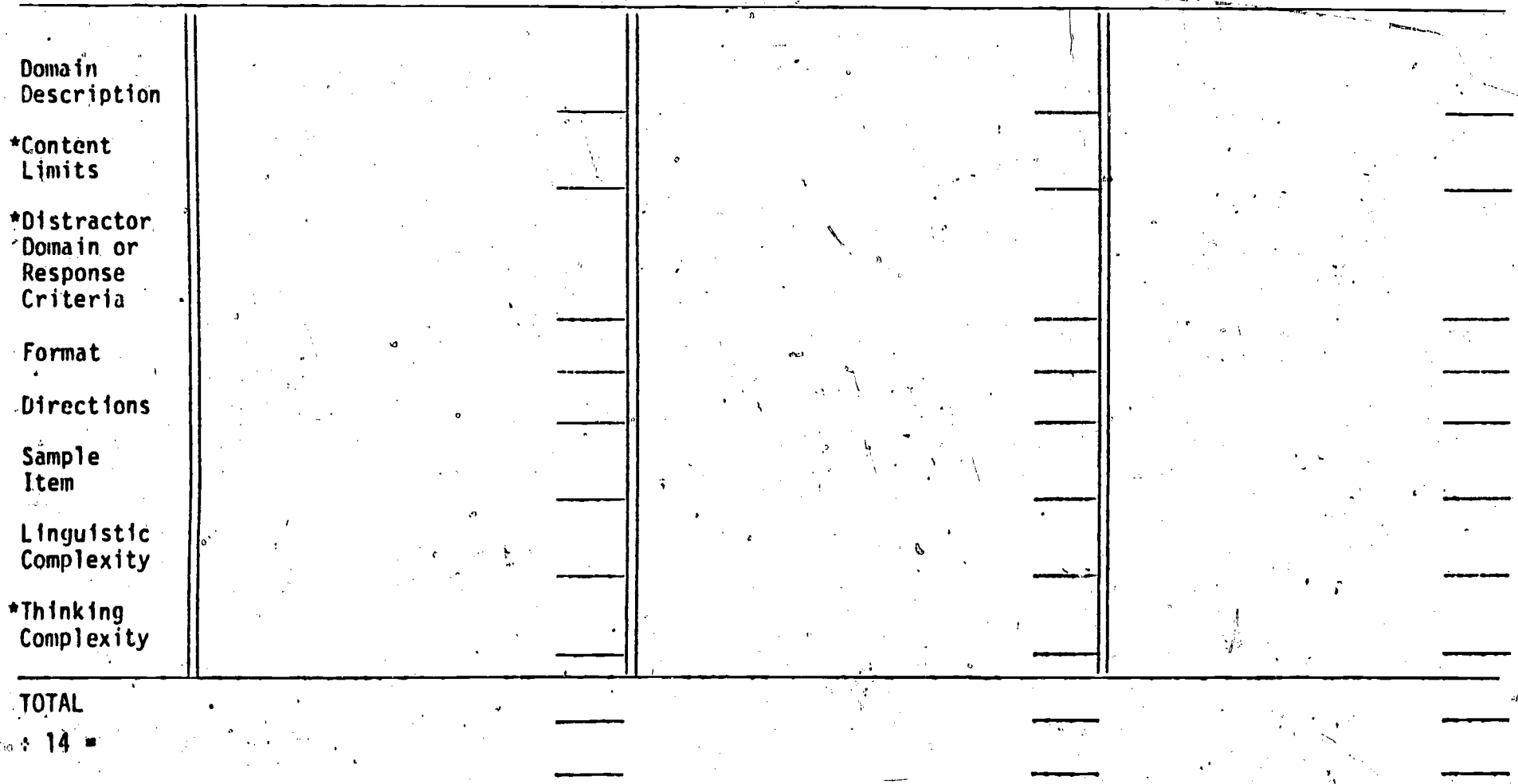
Item Rating Form

SPECIFICATION BEING RATED _____

RATER TITLE _____

COMMENTS: (additional comments can be made on the reverse side)

RATING SCALE



*Critical features

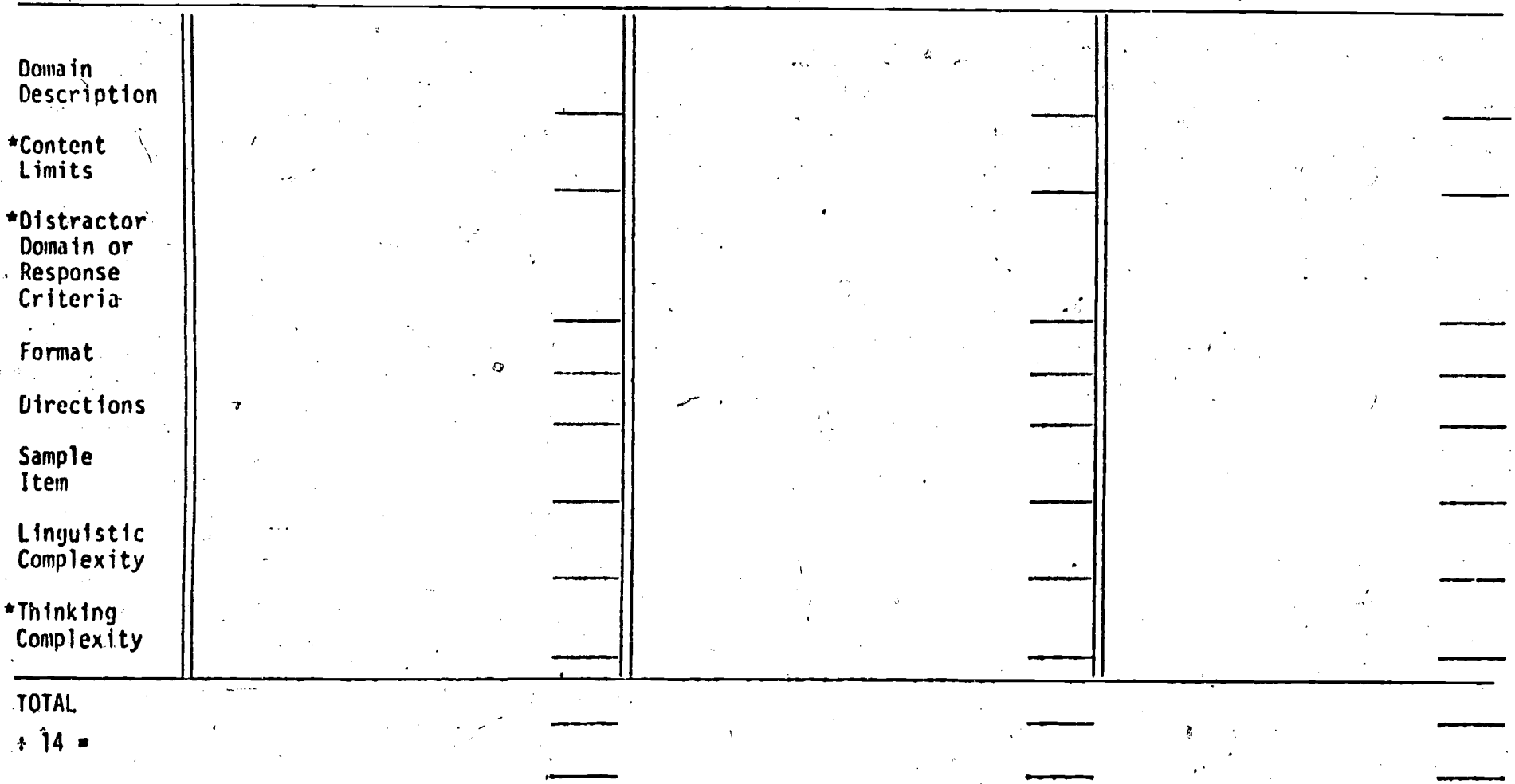
2.18

SPECIFICATION BEING RATED _____

RATER TITLE _____

COMMENTS: (additional comments can be made on the reverse side)

RATING SCALE



*Critical features

2.19

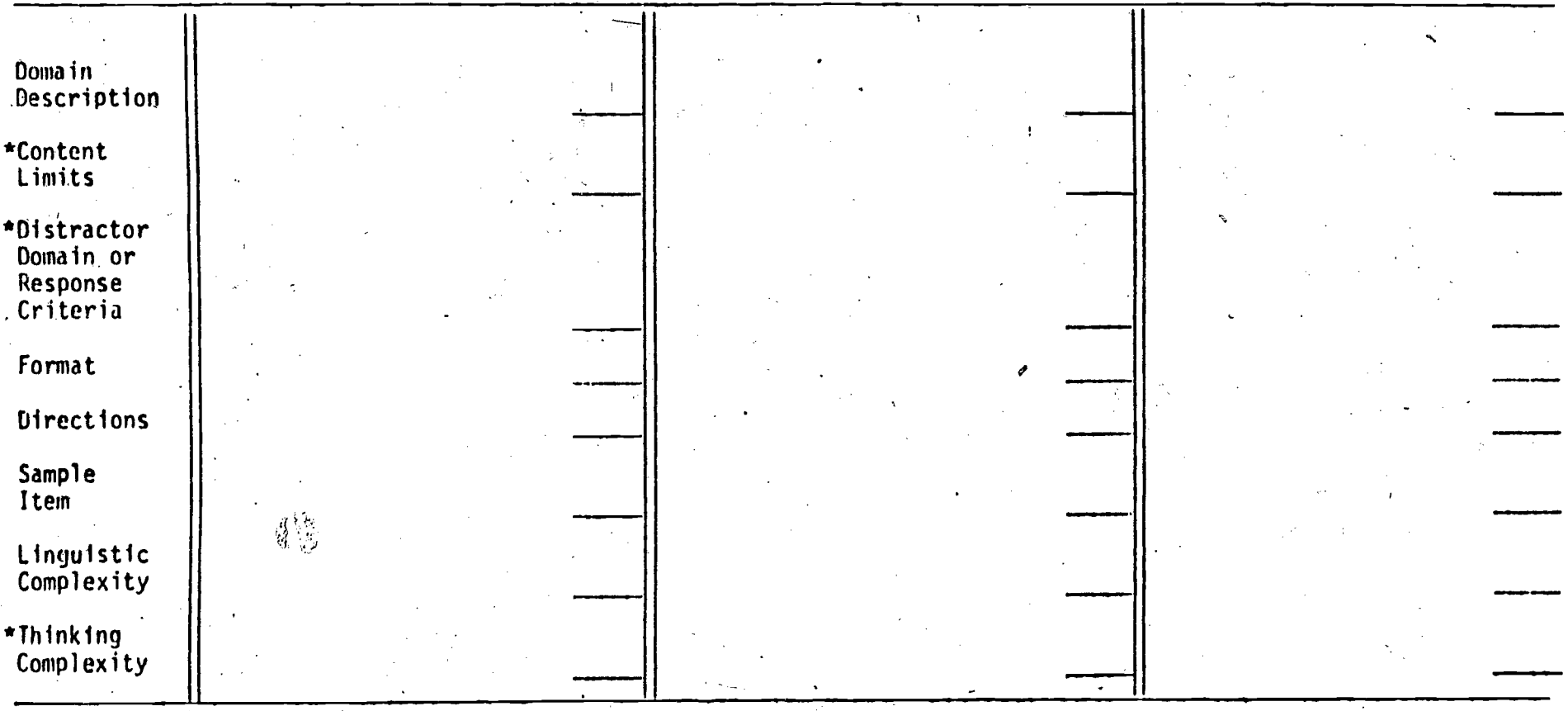
Item Rating Form

SPECIFICATION BEING RATED _____

RATER TITLE _____

COMMENTS: (additional comments can be made on the reverse side)

RATING SCALE



TOTAL

÷ 14 =

2.20

*Critical features

ELEMENTARY MATHEMATICS-SET THEORY

Grade Level: Grade 5

Subject: Elementary Mathematics-Set Theory

Domain Description: Recognition and application of the set theory concepts of membership, subset, intersection, and union for numeric and non-numeric sample sets for simple open sentences involving arithmetic at or below the third grade level.

Content Limits:

1. For multiple choice questions on membership, stimuli may include: a description of membership rules governing a set and an array of four elements, only one of which either does or does not belong to that set; or, an array of four sets, only one of which is the set described.

2. For multiple choice questions on subset, stimuli may include: a set and an array of four sets only one of which is a subset or only one of which is not a subset of the original set; or, a description of rules governing set membership and an array of four possible subsets, only one of which is or is not a subset of the described set.

3. For multiple choice questions on union, stimuli may include a pair of sets and an array of four sets only one of which shows the union of the given pair of sets; or, the reverse, i.e., a given union set and an array of four pairs of sets, such that the union of only one of these pairs would result in the given set.

4. For multiple choice questions on intersection, stimuli may include a pair of sets and an array of four sets only one of which is the intersection of the given pair; or, the reverse, a given intersection set and an array of four pairs of sets, such that the intersection of only one pair would result in the given set.

5. Descriptions of membership rules governing set membership may include common words and phrases relating to objects, principles and ideas that are understood by the average forth grade students.

6. Descriptions of membership rules governing set membership may also include solution sets of simple number sentences, as long as discrimination of the correct answer relies upon knowledge of set theory, and not basic mathematical ability or knowledge above the third grade level.

7. The following symbols may be used without a key { } < > ϕ.

The following symbols may not be used unless a key is provided: $\cup \cap \subset \in$.

8. Items are to be written below fifth grade level of readability.

Distractor Domain:

1. For multiple choice questions on membership, distractors may be drawn from response alternatives that are partially or totally incompatible with the given set descriptions.

2. For multiple choice questions on subset, distractors may be drawn from those wrong answers resulting from reversing the set-subset relationship, from mistaking partial subsets (sets with some elements in common but not all) for subsets, or for mistaking union sets with subsets.

3. For multiple choice questions on union and intersections, distractors may be selected from those wrong answers resulting from confusing union, intersection and subset.

4. Unrelated sets may be used as a distractor for no more than one of the response choices.

5. In items using solution sets, answers resulting from anticipated calculation errors are not eligible as distractors.

6. Distractors may not be such that discrimination of the correct answer relies upon student reading comprehension or student knowledge of other subject matter.

Directions: Select the correct answer.

Format: The question will be multiple choice with four alternatives, three distractors and the correct response.

Italics or boldface must be used to highlight the following words: subset, intersection, union. Also, in cases where students are required to select negative examples, the word "not" should also be highlighted.

The words member and element may be used without explanation.

Sample Item: Selection the correct answer.

_____ is NOT SUBSET of {vegetables}?

a. {potatoes, tomatoes, carrots}
b. {vegetables and fruits}
c. {vegetables that are green}
d. {squash}

DIRECTIONS: Circle the letter of the correct answer to each problem below.

1. Find the SUBSET of {yellow, green, blue, red}

   a) {all the colors in the rainbow}

   b) {blue, green, red, yellow, orange}

   c) {yellow, green, brown}

   ✓ d) {yellow, red}

2. ∩ means INTERSECTION. Find the INTERSECTION:
   {Roger, Rick, Ruth, Roberta} ∩ {girls' names} =

   a) {Roger, Rick, Randy}

   b) {∅}

   ✓ c) {Ruth, Roberta}

   d) {Ruth, Roberta, Rachael, Renee}

3. ∪ means UNION. Find the UNION;
   {3,6,9,12,15,18,21} ∪ {2,4,6,8,10,12,14,16} =

   ✓ a) {2,3,4,6,8,9,10,12,14,15,16,18,21}

   b) {6,12}

   c) {all odd numbers less than 21}

   d) {all even numbers less than 21}

4. John is thinking of a set of numbers whose members fit this number sentence:

$$X + 4 > 12$$

Find the set that is NOT a SUBSET of John's set {X}

✓ a) {all numbers > 8}

b) {9}

c) {14,20,36}

d) {12}

5. Ann is thinking of a set of numbers whose members fit this number sentence:

$$N - 7 < 36$$

Find the set that is NOT a SUBSET of Ann's set {N}

a) {42}

✓ b) {42, 43, 44}

c) {all even numbers < 43}

d) {all odd numbers < 43}

6. $\subset$ means SUBSET. Which of these pairs of sets shows a SUBSET?

   a) {dogs}$\subset${collie, shepard, beagle}

   b) {dogs, cats, birds, fish}$\subset${Spot, Fido, Fluffy, Polly, Goldie}

   c) {dogs, hamsters, guinea pigs, horses}$\subset${pets that live in cages}

   ✓ d) {dogs, cats, hamsters}$\subset${pets}

7. $\cap$ means INTERSECTION. Find the INTERSECTION of set A and set B

   A = {Sam, Steve, Stuart, Sandy} B = {Sue, Sally, Sarah, Sandy}

   $A \cap B =$

   a) {names beginning with S}

   b) {Sam, Steve, Stuart, Sandy, Sue, Sally, Sarah}

   c) {Sam, Steve, Stuart, Sue, Sally, Sarah}

   ✓ d) {Sandy}

ELEMENTARY SCIENCE-GEOLOGY

Grade Level: Grade 7 or 8

Subject: Elementary Science-Geology

Domain Description: Recognizing cause-effect relationships of destructional forces and constructional forces that alter the surface of the earth.

Content Limits:

1. Constructional forces include the following:

volcano:	pressure forces magma (lava) to break through the earth's crust
folding:	forces press the earth's crusts sideways, causing rock layers to become folded upward
earthquake/ faults:	settling and shaking down the earth's crust

Destructional forces include the following:

erosion:	flowing water bumping and wearing away the rock and land, pulling away pebbles and boulders that hammer away at the land as they travel
	wind erosion, sand storm blasting and wearing away the surface of the land
glacier action:	scrape and drag ice and rock across the surface of the land deepening valleys and smoothing out the rocky mountains and hills
lichens:	break up rocks by acid secretions
sunlight/ freezing:	cracks--expansion and contraction of rocks causes break-up

2. Pictorial representations of causes or effects may be used if labelled and accompanied by a verbal promp telling the given part of the item.

3. All prose material should be at or below grade 7 readability.

Distractor Domain:

1. Distractors must be the result of mismatches (mixups) between causes and effects, or misidentification of causes or effects (misnaming).

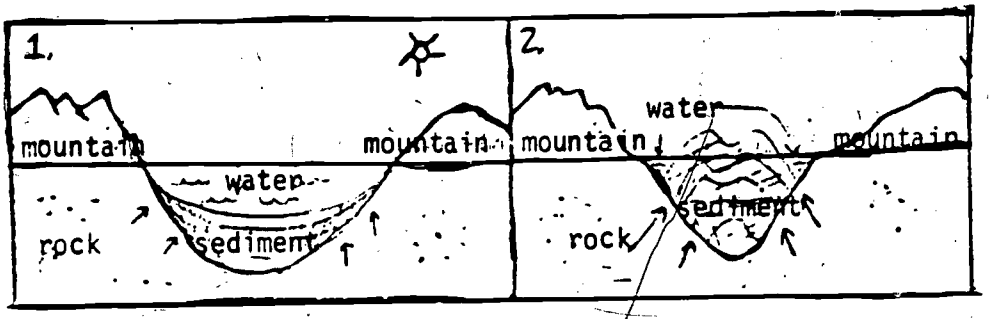
2. Distractors must not be from outside the content limits.

3. Distractors must not be the result of inability to decipher the meaning of pictorial representations.

Format: The questions will each be multiple choice with four response alternatives, three distractors and the correct response.

Directions: Students will be asked to select the correct answer. Each item will pose its own specific question.

Sample Item:



Pressure from surrounding rock pushes the sediment upward, creating mountains. This process is called_____.

a. erosion
b. faulting
c. erupting
✓ d. folding

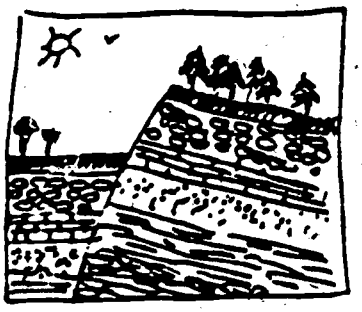
Circle the letter of the correct answer for question 1-4

1. Lichens are _____.

   a) tall, block-shaped mountains formed by faults
   b) tiny organisms that form sediment deposits in river beds.
✓ c) plants that grow on rocks and secrete an acid from the roots.

2. Solar energy can be conceived of as a destructive force in modifying the face of the earth in that the thermal effects upon rocks is to expand them; while in the absence of solar heat, cold temperatures affect contraction. This expansion and contraction cycle _____.

   a) causes fault lines to deepen and widen creating cracks in the earth's surface.
   b) exerts pressure upon the magma below the earth's crust, sometimes leading to volcanic eruption.
✓ c) affects cracks and, eventually, promotes break-up of large boulders and rock surfaces.
   d) erodes the topsoil, allowing the winds to transform valuable farmlands into veritable wastelands.

3. 

   In this picture, layers of the earth's crust have tilted and shifted, creating _____.

   a) a volcano.
✓ b) a fault.
   c) rocks.
   d) a glacier.
   e) a landslide

4. Wind and rushing water can cause _____.

   a) high tides.
   b) cracks in rocks.
   c) pressure on sediments.
✓ d) erosion.

Grade Level: Grade 7 and/or 8

Subject: Elementary science-geology

Domain Description: Applying knowledge of destructional forces and constructional forces that alter the earth's surface, to make predictions of effects, given causes, and/or to hypothesize the causes of given effects.

Content Limits:

1. Constructional forces include the following:

volcano:	pressure forces magma (lava) to break through the earth's crust
folding:	forces press the earth's sediments sideways, causing rock layers to become folded upward
earthquake/ faults:	settling and shaking down the earth's crust

Destructional forces include the following:

erosion:	flowing water bumping and wearing away the rock and land, pulling away pebbles and boulders that hammer away at the land as they travel
	wind erosion, sand storm blasting and wearing away the surface of the land
glacier action:	scrape and drag ice and rock across the surface of the land deepening valleys and smoothing out rocky mountains and hills
lichens:	break up rocks by acid secretions
sunlight/ freezing:	cracks--expansion and contraction of rocks causes break-up

2. Items should not use key terms, e.g., erosion, to pose the problem, but should use description or definitions to convey the process, function.

3. Pictorial representations of causes or effects may be used if labelled and accompanied by a verbal prompt telling the given part of the item.

4. All prose material below grade 7 readability.

Distractor Domain:

1. Distractors must be the result of mismatches (mixups) between causes and effects, or misidentification of causes or effects (misnaming).

2. Distractors must not be from outside the content limits.

3. Distractors must not be the result of inability to decipher the meaning of pictorial representations.

Format: The questions will each be multiple choice with four response alternatives, three distractors and the correct response.

Directions: Students will be asked to select the correct answer. Each item will pose its own specific question.

Sample Item: Flowing water, such as a river, can change the surface of the land by _____.

a) creating a strong pressure against sediments forcing them upward into mountains.
b) giving off acids which slowly eat away at the rocks making them crumble apart.
c) causing sudden shifts upward or downward of great rock masses and layers.
✓ d) carrying pebbles and rocks that scratch and hammer away at the land and rock surface.

Test Items

Circle the letter of the correct answer for questions 1-3.

1. The Dust Bowl was caused by _____.

   a) wind
✓ b) erosion
   c) floods
   d) glacier

2. Glaciers that once existed in North America are responsible for changing of the surface of the earth by _____.

✓ a) scraping and dragging ice and rock across the land.
   b) flooding the land with melted ice and snow.
   c) erupting and spreading lava (magma) over the land.
   d) settling and shaking the layers of the earth's crust.

3. Destruction of the earth's surface can be caused by _____.

   a) faults
✓ b) sun
   c) volcanos
   d) earthquakes

4. Match each cause with its effects.

CAUSES	EFFECTS
_____ faults	a) acid causes rocks to crumble and break up
_____ flowing water	b) deepens valleys and smooths out rocky mountains and hills
_____ sunlight & freezing temperatures	c) expansion and contraction causes rocks to crack
_____ glacier action	d) washes away pebbles and soil, causing erosion
	e) magma breaks through to the surface as lava

INTRODUCTION TO TEST SELECTION - Module III: Comparing Tests' Relevance to a Given Curriculum

This module is concerned with comparing tests' curricular relevance. The procedures involve a series of judgments about curriculum objectives and test materials, expressing these judgments as numbers, combining the numbers for a single test, then comparing the results across tests.

Because the method is a detailed one, it is probably best to use 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 the 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 the curriculum 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 a given curriculum may also vary across levels.

The methods in this module ask you to compare test items with curriculum skills. There are several reasons for carrying out such a thorough analysis of tests before choosing one. First, the analysis helps you to find the

test that is most responsive to your needs; many tests are likely not to match your curriculum 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 skills covered, and the appropriateness of the test's difficulty for the intended students. Finally, the process of assigning numbers to each stage of judgment and carrying them to the next stage ensures that information from earlier judgments is not forgotten or lost. In other words, 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. The best people to do the job would need to be very familiar with the curriculum at the relevant level, have some skill in writing and recognizing objectives, and believe in the importance of curricular relevance in tests.

The Importance of Curricular Relevance in Tests

An extremely important feature to consider in test selection 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 curriculum. Tests of

skills not taught by the local curriculum 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* demonstrated that the content of certain standardized tests is not very standard. The authors found that norm-referenced tests of reading achievement reflect the vocabulary of different basal reading series unequally. That is, a given test 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 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.** 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 balance between exercises on literal versus inferential comprehension was 83% to 17%, but for the other two series it was about 42% to 58%. Two types of comprehension skills--cloze sentences

---

* Jenkins and Pany, 1976.

** Armbruster, et al, 1977.

and words in context--were covered in one or more reading series, but were not included in either test. Cloze sentences made up 24% of the comprehension exercises in one reading series, 51% in the second series, and 28% in the third. The words-in-context represented 1%, 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,* the authors found that four widely used norm-referenced 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** 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 such broad skills. Coverage of the reading objectives by eight popular norm-referenced test 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 these 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

---

* Porter, et al, 1978.

** Hoepfner, 1978.

the present purposes, the relevant result 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 (viz., 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 so 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 of students' performance on tests. One further study* 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 this conclusion:

> What these studies show, apparently, is not that the new 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)

---

* Walker and Schaffarzik, 1974.

The first four studies show that the content of standardized tests differs and that these tests differ in their correspondence with any given curriculum. The conclusion that such variation in test content could, irrespective of students' actual achievement, bias the outcomes is confirmed by the fourth 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 criterion-referenced tests are not all the same, curricular relevance may be equally a problem for criterion-referenced tests and for norm-referenced tests. In contrast with norm-referenced tests, however, criterion-referenced tests give a separate score for each objective, thus making it easier to distinguish students' performance on curriculum-relevant and curriculum-irrelevant objectives.

Checklist Step 1:

Prepare a listing of the objectives of the curriculum component to be tested.

To find the test most relevant and responsive to your curriculum, it is necessary to be very clear about the instructional objectives 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 is a discrepancy between the operational classroom curriculum and the official, formal one. Or you may be confronted with a situation in which 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.

Suggestions are given here for drawing up your list of curricular objectives under two conditions:

- When each subject area to be tested 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. <u>When there is a uniform curriculum, list or index the objectives for the program component to be tested as follows:</u>

(1) Write the objectives in enough detail so that later in the process it will be possible to judge how closely a given test item measures or matches an objective. If, for example, your math course teaches division in working (i.e., radical) form, but a test formats its division problems in number sentence form, your listing of local math objectives should enable the test taker to detect this difference and judge its importance. Similarly, 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 curriculum. Since formal curricula are often stated in rather general or global objectives, 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 teaching 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 strictly verbal 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 headings

like geometry, measurement, money, time, graphing, word problems, basic operations, and the like. For elementary reading, objectives could be grouped under headings like 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 actually covers. The objectives in Column 1 of the sample CSE Test Relevance Rating Form are grouped under headings labeled Curricular Subareas and Skill Clusters.

(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 by reduced to six objectives dealing with consonants and vowels in each of the three positions. These six larger 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 larger skills which the small skills comprise.

Two cautions should be noted about combining objectives. First, the amount of combining that is useful will vary with the intended use of the test. More combining will be useful 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), then the descriptions or labels for those objectives do not make it clear what is being taught, learned, or tested. Such broad spectrum objectives do not describe the 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 objectives. If, however, the formal curriculum accurately represents current 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. When the operational, classroom curriculum varies, but there is great commonality in the important objective for the component to be tested, make a listing of the common objectives as follows:

(1) Either compare listings of the separate classroom curricula and make a 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 check off the objectives which they actually teach at that level. Make

a single curriculum-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.

Checklist Steps 2 and 3:

Write your listing of curriculum objectives to be tested in Column 1 of the Test Relevance Rating Form, and then record the number of objectives in Box B.

Column 1 of the worksheet will contain your listing (or indexing) of the curricular component to be tested. The total listing will be organized so that related objectives are grouped together under a common heading. Some of the smaller, more detailed objectives in your curriculum may not appear separately in the listing because they have been grouped together into one larger objective.

Several sheets may be needed for listing or indexing the component to be tested. Number the pages and draw a heavy line under the last 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. On the rating form for the module exercise there are 10 curriculum objectives listed. A sample of a completed worksheet is also included in this packet of materials.

Checklist Steps 4 and 5:

Rate the importance of objectives. Then duplicate the worksheet and fill in the identifying information for each test to be rated.

In this step, judgments are made about the importance of each of the objectives that are listed in Column 1. These judgments are then expressed as numbers, indicating degrees of importance, and are later recorded in the third column.

For each of the 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 necessary for student progress, are given a value of 3.

After judging the importance of each 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 the ratings seem appropriate relative to each other.

On completing all of the steps up to this point, make enough copies of the partially filled-in 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.

Checklist Steps 6 and 7:

List/index all of the items on the test in Column 2 of the Test Relevance Rating Form, each on the same line as the curriculum objective that is most closely related to it.

Look at each test item and decide which 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 skill. At this stage, be generous in judging whether an item is responsive to an objective; what is important here is to assemble with each objective all of the items that measure it, even remotely.

Try to pair each test item with only one curriculum objective; but if an item seems to measure more than one skill, write its number in Column 2 opposite each skill. 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 in Column 2 next to the relevant objectives. In fact the objectives of many 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 that number as the denominator in Box A and C on the final page 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 means do not count any circled (i.e., repeated) items.

Checklist Step 8:

Judge how closely the test items correspond with the respective curriculum skills, and record these judgments in Column 4 of the Test Relevance Rating Form.

The purpose of this step is to judge how relevant or sensitive each item is to the corresponding skill that your curriculum teaches. Examine each test item, and judge how closely it corresponds to the respective objective in format, content, and skill tested. The correspondence may be unacceptable, adequate, or very close. For those degrees of match/mismatch, assign a score of 0, 1, or 2 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 skills on the test. If the answer is yes, than a rating of 2 is not appropriate for that item. If the test format is so unfamiliar as to make it very hard for students to show their learning of the skill, then a zero rating should be recorded.

Attend also to the content and process that 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.

In the module's exercise, the issue of curriculum and item content is illustrated by comparing the first skill with the respective test items. The objective calls for specific affixes and also for some words which do not have affixes. For such judgments you may need to set some arbitrary criteria, such as these:

90 - 100% Congruence rates a 2

80 - 90% Congruence rates a 1

<80% Congruence rates a 0 as unacceptable

The issues of item format and item solution processes are illustrated by comparing the skill on compound words with two sets of items on the sample test, #'s 7-9 and 10-14 (p. 3.40). The objective calls for a matching format involving two column of real words; so do items #'s 10-14. But items #'s 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 curriculum, and probably much less familiar.

Item format usually affects the mental processes which a pupil must use for coming up with correct answers. In items #'s 7-9, pupils need to be able to understand the concept "all four words" and to keep in mind while reading the words. The test-takers also need to break down each word in #'s 7-9, sometimes more than once, e.g.,

fi - replace

fire - place

and judge each part for whether it is a real word. 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 on a line 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 on the line are compounds, the test-taker circles Yes and goes on.

In contrast, the processes for solving #'s 10-14 involve remembering a word on the left, building possible compounds out of it and words on the right, judging each possible compound, and continuing until a compound is recognized.

If the differences between the curriculum skill and the content/format/process of item #'s 7-9 will interfere with your pupils using their 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 skill, or unacceptable. Record the rating for each of the items in the column headed Step 8 on the line where the respective items are indexed.

A second example of a difference between a curriculum skill and a tested one occurs with items #'s 33-35 (p. 3.45) on inferential comprehension. The skill asks for stories which are about three paragraphs long. The items use a text which is rather short. If you think that that difference does not really change the skill, 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 curriculum and test text length does change the skill somewhat, then assign and record a lower congruency rating.

Checklist Step 9:

Rate the appropriateness of the difficulty of each test item and record the ratings in the fifth column of the Test Relevance Rating Form.

The last judgment of test materials 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 objective will answer it correctly much of the time?
- Is the item so difficult that students who have mastered the skill 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. If 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 the curriculum 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.

Checklist Steps 10 and 11:

For each objective that has any acceptable test items, multiply the ratings for each item (Column 3 x Column 4 x Column 5) and record the products in Column 6 of the worksheet. Then find the sum of these products.

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 not precise measures; they are summaries of the test rater's judgments about the importance, curricular relevance, and difficulty of the skills 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 skill (rated 3 in Column 3)
- Matches the skill 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 on the column. Also, write it in Box A as the numerator.

Checklist Step 12:

Count the number of acceptable items on the test and write it in Box C of the worksheet as the numerator.

As a step toward finding the proportion of the test's items which are relevant to your curriculum, 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.

Checklist Steps 13 and 14:

Compute 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 has little meaning. It 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 curriculum skills. This one comparison still does not indicate whether the highest rated test covers the curriculum well enough. That judgment is aided by two other summary figures on the worksheet, the Index of Coverage and the Index of Relevance.

The Index of Coverage tells how completely a test covers the curriculum 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 objectives 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 curriculum skill, 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 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 skills 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.

The last summary figure for comparing tests is the Index of Relevance, which tells what proportion of the test is sufficiently relevant to your curriculum. It is computed by dividing the total number of items on the test (Step 7) into the number of items that adequately match the curriculum (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 value ranging from zero (totally unresponsive to the local curriculum) to 1.0 (all of the test items are adequate measures of curriculum skills). On a test with a relevance rating of .75, a quarter of the items measure skills 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 curriculum means paying, both in time and money, for test materials that work against you. Your students may do poorly on skills in the test which do not match your

curriculum, and the test results will not be very helpful for assigning lessons.

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 Test Relevance Rating Form. For example, the average importance of curriculum skills 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 curriculum-irrelevant can later be omitted from the analysis of scores. Total test scores could be reported, if required by higher authority, but the customized, curriculum-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 unit 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 into a summary rating of a test 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. The exercise materials in this module can be used for training, or else a part of your curriculum may be used, for familiarization, with a real test. Next, it will help to have some discussion of the judgmental scales to encourage 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 the 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.

Although the procedures in this unit 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.

REFERENCES

Armbruster, B.B., Stevens, R.J., & Rosenshine, B. Analyzing content coverage and emphasis: a study of three curricula and two tests. Technical Report #26, Center for the Study of Reading, University of Illinois, Urbana, Illinois, 1977. ERIC # ED 136 238.

Hoepfner, R. Achievement test selection for program evaluation. In Wargo, M.J., & Green, D.R. (Eds.), Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. CTB/McGraw-Hill, 1978.

Jenkins, J.R., & Pany, D. Curriculum biases in reading achievement tests. Technical Report #16, Center for the Study of Reading, University of Illinois, Urbana, Illinois, 1976. ERIC # ED 134 938.

Porter, A., & Floden, D. Don't they all measure the same thing? Paper given at the Conference on Measurement and Methodology, Center for the Study of Evaluation, Los Angeles, 1978.

Walker, D.F., & Schaffarzik, J. Comparing curricula. Review of Educational Research, 1974, 44, 83-112.

HOW TO SELECT A TEST

Comparing tests' relevance to a given curriculum

Checklist

____ 1. Prepare a listing of the part of the curriculum that you want to test.

____ 2. Enter your listing of objectives in the first column of the Test Relevance Rating Form, called the worksheet.

____ 3. Count the number of program skills and enter in Box B on the final page of the worksheet.

____ 4. Rate the importance of each objective in your listing and record these judgments in the third column of the worksheet.

____ 5. Make enough copies of the worksheet for all of the raters and all of the tests still under consideration. Then for each of the tests fill in the blanks at the top of the pages of the worksheet.

____ 6. For each test, index its items in Column 2 of the worksheet next to the objectives they relate to.

____ 7. Count the number of items on the test and enter in both Box A and Box C on the final page of the worksheet.

____ 8. Judge how closely a test's items correspond to the respective program skills in format, content, and process. Record these judgments in the fourth column of the worksheet.

____ 9. Rate the appropriateness of the difficulty of the test items, and record these ratings in the fifth column of the worksheet.

____10. For each program objective that has any items on the test, multiply the ratings in Columns 3, 4, and 5, and enter the products in the sixth column of the worksheet.

____11. Add all of the products from Step 10 and record at the bottom of the sixth column and in Box A of the worksheet.

____12. Count the number of adequate test items (i.e., the number of numbers in Column 6) and record as the numerator in Box C.

____13. Compute the summary indices of tests' congruence with the curriculum (namely, the Grand Average, the Index of Relevance, and the Index of Coverage) and enter at bottom of last page and in spaces at top of front sheet of the worksheet.

____14. Compare the summary indices of the tests under consideration. Decide whether one test has markedly greater congruence with your curriculum.

CSE TEST RELEVANCE RATING FORM (partially filled in for Workshop Exercise)

Step 5

Test name, level, and form All American Test of Reading Comprehension, brown level Rated by Marion Choy

Program subject and level 5th/6th grade reading comprehension Date 1/15/76

Overall ratings *(fill in last)*: Grand Average _____ *(average congruence per item ranging from 0-6)* Index of Coverage _____ *(proportion of program skills measured by test)* Index of Relevance _____ *(proportion of the test that is relevant to your program skills)*

Step 2 Listing of Program Skills	*Step 6* Index of corresponding test items	*Step 4* Importance of program skills *1=minor 2=important 3=essential*	*Step 8* Match between items and skills *0=not acceptable 1=adequate 2=very close*	*Step 9* Appropriateness of item difficulty *0=too hard or too easy 1=acceptable*	*Step 10* Combined judgments *Products across columns 3, 4, 5*	Notes
Word level objectives						
Curricular Sub-area → Word attack Skill cluster →						
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. 3.39 #1 2 3 4 5 6					
Compound words: Pupils will complete compound words by matching words in a left column with words in a right column.	p 3.40 #7 8 9 p 3.40 #10 11 12 13 14					

111

112

CSE TEST RELEVANCE RATING FORM

Test name, level, form All American Test of Reading Comprehension, brown level Rater Choy Date 1/15/76

Program Skills	Index of items	Importance of program skills	Match between program skill & item	Appropriateness of item difficulty	Products of prev. 3 columns	Notes
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 3.41 #15 16 17	1				
Meaning						
Synonyms: Given a vocabulary word, the pupil will select from multiple choices the word or phrase which is a synonym.	p 3.42 #18 19 20					
Antonyms: Given a vocabulary word which has an opposite, the pupil will select its antonym from multiple choices.	p 3.42 #21 22 23					

CSE TEST\RELEVANCE RATING FORM

Test name, level, form All American Test of Reading Comprehension, brown level Rater Choy Date 1/15/76

Program Skills	Index of items	Importance of program skills	Match between program skill & item	Appropriate-ness of item difficulty	Products of prev. 3 columns	Notes
Phrase, sentence, and text level objectives						
Meaning from context - words with one familiar meaning: Given sentences with one 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.	p 3.43 #24 25 26					
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 3.43 #27 28 29					
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 3.44 #30 31 32					

3.30

Final sheet
page 4

CSE TEST RELEVANCE RATING FORM

Test name, level, form All American Test of Reading Comprehension, brown level Rater Choy Date 1/15/76

Program Skills	Index of items	Importance of program skills	Match between program skill & item	Appropriateness of item difficulty	Products of prev. 3 columns	Notes
Inferences: Given a story in about three paragraphs, pupils will mark whether each of several supposed inferences from the story is probably true, probably false, or can't tell.	p 3.45 #33 34 35					
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.3.46 #39 40 41				*Step 11* (Sum of numbers in sixth column. Write it in Box A also.)	
END OF LISTING	p 3.45 #36, 37, 38 Clearly irrelevant items					

118

OVERALL RATINGS *Step 13*

117

Box A

GRAND AVERAGE: _____

Sum of numbers in 6th column *(step 11)* _____

divided by _____

Total number of test items *(step 7)* _____

Box B

INDEX OF COVERAGE: _____

Number of program skills adequately measured by test _____

divided by _____

Total number of program skills in first column *(step 3)* _____

Box C

INDEX OF RELEVANCE: _____

Number of acceptable test items *(step 12)* _____

divided by _____

Total number of items on the test *(step 7)* _____

3.31

4 sheets

CSE TEST RELEVANCE RATING FORM (worked example)

Test name, level, and form All American Test of Reading Comprehension, brown level Rated by Marion Choy

Program subject and level 5th/6th grade reading comprehension Date 1/15/76

Overall ratings* *(fill in last)*: Grand Average 2.1 *(average congruence per item ranging from 0-6)* Index of Coverage .70 *(proportion of program skills measured by test)* Index of Relevance .63 *(proportion of the test that is relevant to your program skills)*

Step 2 Listing of Program Skills	Step 5 Index of corresponding test items	Step 4 Importance of program skills *1=minor 2=important 3=essential*	Step 8 Match between items and skills *0=not acceptable 1=adequate 2=very close*	Step 9 Appropriateness of item difficulty *0=too hard or too easy 1=acceptable*	Step 10 Combined judgments *Products across columns 3, 4, 5*	Notes
Word level objectives						
Curricular Sub-area → Skill cluster: Word attack						
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 3.39 #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	
Compound words: Pupils will complete compound words by matching words in a left column with words in a right column.	p 3.40 #7	1	0			Format is way off.
	8		0			Not similar enough to program skill
	9		0			
	p 3.40 #10		2	1	2	
	11		2	1	2	
	12		2	1	2	
	13		2	1	2	
	14		2	1	2	

113

120

Note: these ratings will vary with your judgments of your pupils' abilities and the importance of the program skills.

CSE TEST RELEVANCE RATING FORM

Test name, level, form All American Test of Reading Comprehension, brown level Rater Choy Date 1/15/76

Program Skills	Index of items	Importance of program skills	Match between program skill & item	Appropriateness of item difficulty	Products of prev. 3 columns	Notes
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 3.41 #15 16 17	1 ↓	0 0 0			More important as a writing skill than a reading one. Item format calls for pupils to select root words from four very different choices.
Meaning						
Synonyms: Given a vocabulary word, the pupil will select from multiple choices the word or phrase which is a synonym.	p 3.42 #18 19 20	2 ↓	2 2 2	1 1 1	4 4 4	
Antonyms: Given a vocabulary word which has an opposite, the pupil will select its antonym from multiple choices.	p 3.42 #21 22 23	2 ↓	2 2 2	1 1 1	4 4 4	

3.33

CSE TEST RELEVANCE RATING FORM

Test name, level, form All American Test of Reading Comprehension, brown level Rater Choy Date 1/15/76

Program Skills	Index of items	Importance of program skills	Match between program skill & item	Appropriateness of item difficulty	Products of prev. 3 columns	Notes
Phrase, sentence, and text level objectives						
Meaning from context - words with one familiar meaning: Given sentences with one 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.	p 3.43 #24	1	2	1	2	
	25	1	2	1	2	
	26	1	2	1	2	
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 3.43 #27	3	1	0		Too hard
	28	3	1	0		
	29	3	1	0		
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 3.44 #30	3	2	1	6	
	31	3	2	1	6	
	32	3	2	1	6	

123

124

3.34

Final sheet
page 4

CSE TEST RELEVANCE RATING FORM

Test name, level, form All American Test of Reading Comprehension, brown level Rater Choy Date 1/15/76

Program Skills	Index of items	Importance of program skills	Match between program skill & item	Appropriateness of item difficulty	Products of prev. 3 columns	Notes
Inferences: Given a story in about three paragraphs, pupils will mark whether each of several supposed inferences from the story is probably true, probably false, or can't tell.	p 3.45#33 34 35	1 1 1	2 2 2	1 1 1	2 2 2	The small difference between item and program paragraph length does not seem important.
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 3.46#39 40 41	1 1 1	2 2 2	0 0 0	Step 11 88 (Sum of numbers in sixth column. Write it in Box A also.)	Too easy. The distractors don't make sense, so they couldn't be correct choices.
END OF LISTING	p 3.45#36, 37, 38 Clearly irrelevant items					

OVERALL RATINGS *Step 13*

Box A

GRAND AVERAGE: 2.1

Sum of numbers in 6th column *(step 11)* 88

divided by

Total number of test items *(step 7)* 41

Box B

INDEX OF COVERAGE: .70

Number of program skills adequately measured by test 7

divided by

Total number of program skills in first column *(step 3)* 10

Box C

INDEX OF RELEVANCE: .63

Number of acceptable test items *(step 12)* 26

divided by

Total number of items on the test *(step 7)* 41

125

126

First sheet of _____ sheets

CSE TEST RELEVANCE RATING FORM

Step 5

Test name, level, and form _____ Rated by _____

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 skills measured by test)* Index of Relevance _____ *(proportion of the test that is relevant to your program skills)*

Step 2 Listing of Program Skills	*Step 6* Index of corresponding test items	*Step 4* Importance of program skills *1=minor 2=important 3=essential*	*Step 8* Match between items and skills *0=not acceptable 1=adequate 2=very close*	*Step 9* Appropriateness of item difficulty *0=too hard or too easy 1=acceptable*	*Step 10* Combined judgments *Products across columns 3, 4, 5*	Notes

123

127

Continuation sheet
page ____

CSE TEST RELEVANCE RATING FORM

Test name, level, form _____ Rater _____ Date _____

Program Skills	Index of items	Importance of program skills	Match between program skill & item	Appropriateness of item difficulty	Products of prev. 3 columns	Notes

3.37

Final sheet
page _____

CSE TEST RELEVANCE RATING FORM

Test name, level, form _____ Rater _____ Date _____

Program Skills	Index of items	Importance of program skills	Match between program skill & item	Appropriate-ness of item difficulty	Products of prev. 3 columns	Notes
	Clearly irrelevant items				*Step 11* _____ (Sum of numbers in sixth column. Write it 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 6th column *(step 11)* _____	Number of program skills adequately measured by test _____	Number of acceptable test items *(step 12)* _____
divided by _____	divided by _____	divided by _____
Total number of test items *(step 7)* _____	Total number of program skills in first column *(step 3)* _____	Total number of items on the test *(step 7)* _____

All-American Test of English
Reading Comprehension

Brown Level

Some of these materials are adapted from the item banks of Downers Grove, Illinois, Unified School District and the El Dorado, California, County-School District.

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:

re<u></u>write
happy
watch<u></u>ful

Draw a line under each prefix or suffix.

1. dislike
2. during
3. driver
4. people
5. quickly
6. refill

Go to the next page

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:

eye	fruit
grape	knob
door	brow

Explanation
eyebrow
grapefruit
doorknob

10-14.

an	noon
after	fly
any	light
butter	body
flash	other

Go to the next page

DIRECTIONS: In each problem below put a check next to the one word that is a root word. The first problem is done for you.

EXAMPLE: Check the one word that is a root word:

✓ paper

____ selfish

____ naughty

____ unsafe

15. Check the one word that is a root word:

____ looked

____ happily

____ dirty

____ cold

16. Check the one word that is a root word:

____ thirsty

____ family

____ talking

____ smiled

17. Check the one word that is a root word:

____ arrange

____ filled

____ books

____ winning

Go to the next page

DIRECTIONS: Read the word on the left below. Then circle the one word on the same line that has the same meaning. The first problem is done for you.

EXAMPLE:	flat:	rough	tall	(level)
18.	blonde:	hairy	fair-headed	brunette
19.	wide:	narrow	broad	long
20.	steal:	iron	give	rob

DIRECTIONS: Circle the word that means the opposite of the word on the left. The first problem is worked for you.

EXAMPLE:	near:	high	(far)	away
21.	warm:	hot	cold	cool
22.	happy:	sad	silly	funny
23.	rough:	soft	smooth	hard

Go to the next page

DIRECTIONS: Read the first part of each sentence, then circle the word that completes the sentence best. The first problem is done for you.

EXAMPLE: The dog gnawed the _____.

a. boy
b. bone
c. boat

24. To hit the ball the boy needed a _____.

a. glove
b. bat
c. belt

25. The bird ate three _____.

a. words
b. worms
c. nests

26. Mary bought an apple at the _____.

a. barn
b. start
c. store

DIRECTIONS: Read the first sentence, thinking carefully of the underlined word. Then read the other sentences and check the one in which the underlined word has a new or different meaning. The first question is done for you.

EXAMPLE: Mother cut her hand on a can.

a. Which is your right hand?
✓b. Please hand me that book.
c. Put your hand on your head.

27. Susan reads as well as Jane.

a. Jack and Jill went to a well.
b. How well can you tell time?
c. Mike can not draw animals well.

28. Tom can't find his pen.

a. This pen has red ink in it.
b. Dick put the pigs in the pen.
c. Pete wants a pen for his birthday.

29. Do you have anything to eat?

a. Pete will have a birthday party.
b. Have you got any pennies?
c. We have to go home now.

Go to the next page

DIRECTIONS: Read each story below and then check the main idea for the story. The first question is done for you.

EXAMPLE: Gold is soft, almost as soft as putty. It can be hammered into a thin wafer five millionths of an inch thick without being heated. Just one ounce of gold can be beaten into a thin sheet 100 feet square, or drawn into a thin wire stretching fifty miles. In addition, gold is a superb conductor of electricity and a marvelous reflector of heat.

The story mainly tells:

a. Why gold reflects heat and light
b. Why gold is so soft
c. What makes metals so valuable
✓d. What wonderful qualities gold has

30. The frankfurter, named for the city of Frankfurt in Germany, is easily the most popular sausage in the world. Frankfurters, popularly known as "hot dogs," are sold almost everywhere in the United States. They are consumed in great numbers at sporting events and amusement places. People from foreign countries often think hot dogs are one of the characteristics of American life.

The story mainly tells:

a. Why hot dogs are popular
b. How hot dogs and frankfurters differ
c. What foreign people think of hot dogs
d. How popular hot dogs are

31. Why does a mustang buck so wildly when a saddle or man is on its back for this first time? Mustangs have the blood of wild horses. Their ancestors roamed the plains, hunted by wolves and mountain lions. They had a built-in terror of being attacked and killed by fang and claw. Instinctively they became all fear and fire when something leaped on their backs.

The story mainly tells:

a. What the mustangs' ancestors did
b. What animals killed mustangs
c. What makes mustangs buck wildly
d. Why horses are difficult to train

32. A buffalo stampede was a frightening thing to see. The shaggy-headed buffalo, weighing from 1000 to 2000 pounds, rushed blindly forward, bringing death and destruction to anyone and anything unlucky enough to be caught in their path.

The story mainly tells:

a. How heavy buffalo are
b. What a buffalo stampede was like
c. How hard buffalo charge
d. Why people are afraid of some animals

Go to the next page

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 <u>true</u>, check A. If the sentence is <u>probably false</u>, check B. If you <u>can't say</u> 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

DIRECTIONS: Read each sentence and then underline the part of the sentence which shows <u>exaggeration</u>. The first one is done for you.

EXAMPLE: "Don't drop that light bulb, Roger," said Mr. Fairfield. "If you do, it will break into <u>ten million</u> pieces."

36. "Don't walk so heavily, Debbie. It sounds as if an elephant were walking through the hall," scolded Mr. Glass.

37. "This flashlight battery is powerful," said Jerry. "I'll bet it could make a flashlight bright enough to light up all the city at once."

38. "That certainly is a strange-looking animal," said Linda. "Its tail must be a block long. Is it dangerous?"

Go to the next page

DIRECTIONS: Read each sentence and underline the meaning of each colloquial expression. The first one is done for you.

EXAMPLE: When the dog ate the gingerbread, Mrs. Weber was hopping mad.

hopping like an angry rabbit
very angry

39. It is easy to catch cold in very bad weather.

get sick with a cold
catch hold of cold air

40. Our team played very well, and it soon took the lead.

got the better score
grabbed the leading player

41. Mrs. Lane was so worried that she snapped at Ann for no reason at all.

tried to bite
spoke crossly to

END OF TEST:
RAISE YOUR HAND

INTRODUCTION TO TEST SELECTION -- MODULE IV: Comparing the Technical and Practical Merits of Tests

This module discusses characteristics of tests that may be used to compare the tests' overall technical and practical merits.

As an aid in identifying features to use for screening and comparing tests, a listing of 32 such dimensions of tests is included in this module. You are urged to modify the list according to your experience and to the present need for tests. There is almost no end to the number of test features which you could consider, so you may want to add to the list (e.g., a feature such as the availability of in-service training in giving and scoring a test). On the other hand, not all test features are important for a given test use (e.g., alternate test forms are not important to have if you are doing one-shot testing), so you will probably eliminate some of the listed features for any given testing situation.

The listing of suggested test features follows on pages 4.2 to 4.6.

Features that Can Be Used to Screen Tests' Technical and Practical Merits

A. Objectives or domains that a test includes

Sources of information on the feature: Listing in a test manual or continuum chart of the objectives covered at each level of the test

Relevant test uses: Program planning, diagnosis, progress monitoring, proficiency testing

1. Clarity of objectives: Does the statement of each objective or skill make it clear just what content and behavior are being tested; or would many different types of content and behavior be consistent with each objective? Objectives such as the following are not clear: reading comprehension, critical thinking, arithmetic operations, arithmetic applications. The following objective approaches clarity, but still has much leeway:

> Given a story of 4-5 lines at a fourth grade reading level, pupils will select the main idea. The three distractors will deal with particulars of the story or with generalizations from single particulars.

2. Rationale for objectives: The primary justification for including an objective on a test is that the objective is actually taught in the classroom. If a test is being selected to measure the direct effects of instruction, this test feature may be omitted here because it is covered in detail in another module. But if a test is being used for some predictive purpose, such as to test survival skills in reading and math, the choice of objectives should be well justified.

3. Flexibility in selecting objectives: Is an adequate range or number of different objectives covered? For testing at several levels, are the most important core objectives covered at several test levels with items that are appropriate for the respective ages? For continuous progress monitoring (end of unit testing), are single objectives easy to test separately?

4. Number of items per objective: Is the number of test items for each objective appropriate for the intended use of the test? When test results are to be used to make decisions about the specific skills of individual pupils, there must be at least several items per specific skill. For end of unit testing there should be at least eight or ten, and often much more, depending on the desired level of proficiency. When there are only one or two items per objective, then the results may be misleading, because the score for an objective can be greatly affected by carelessness, guessing, or using a neighbor's answer. If a test is to be used to survey a program, and not to support classroom instruction, then the number of items per objective can be small, and the variety of objectives should be great.

B. Adequacy of the test development process

Sources of information on the feature: Technical reports, technical manuals, or technical sections of the test manual

Relevant test uses: all

1. Item review: Were the test items adequately reviewed for clarity, reading level, flaws in item construction, etc., not only by reviewers, but also on the basis of pilot testing?

2. Congruency: Is there convincing evidence that the test items measure the skill and content described by their respective objectives? If there are only sketchy objectives, then there cannot be any convincing evidence. This is important for objectives-based testing.

3. Representativeness: Is each set of items which gets scored a typical, representative sample of the skill? Or is it instead a biased, untypical sample?

4. Pilot testing: Was pilot testing of the test items carried out, and if so, was it on a representative and sizable sample of test takers?

C. Validation by field testing

Evidence: Technical reports or technical manuals

Relevant test uses: All, except where noted.

1. Pupils in the field testing sample: Was the production version of the test validated through field testing, and if so, was the group of students in the field test sizable and either representative of the nation or of the groups to be tested in your program? If not, then any data for C.2-5 below are not meaningful.

2. Sensitivity to learning: Is evidence reported that instruction is followed by dependable increases in scores on relevant items? The evidence should be free of the usual problems in measuring gain (e.g., the increase in test scores should not be attributable merely to maturation of the pupils).

3. Consistency of scores: Is appropriate evidence for the reliability of test scores reported? For tests where human judgment is heavily involved in the scoring it is essential to have evidence of inter-judge or inter-rater reliability. If alternate forms of a test will be used, consistency in their scores for individual test takers should be shown. When scores on individual objectives or subtests will be used (e.g., for diagnosis), then the reliability data should be for such scores, not

merely on total test scores. Finally, if the scale of test scores is to be split into two or three categories (e.g., fail, marginal, pass), the reliability data should be for that type of decision and not for total test scores.

4. Lack of bias in item statistics: Does the test publisher report statistical evidence that the test functions essentially the same for the different pupil groups that you will be testing? A showing that one group performs somewhat lower overall than another is not sufficient evidence, because that result will occur when the one group has not learned the target skills as thoroughly. Sound evidence will consist of data showing the pattern of item difficulties (that is, the difficulty of each item on a test relative to the difficulties of the other items) is the same for the various groups being tested; or that items cluster similarly.

5. Validity of passing scores: If cutting scores are to be used, is evidence offered that pass/fail scores strongly predict a valid indicator of success?

D. Appropriateness for examinees

Sources of information on the feature: Examiner's manual, test materials, answer materials

Relevant test uses: All

1. Surface fairness: Are different racial, national, or cultural groups portrayed, in words or pictures, representatively and positively?

2. Vocabulary of test items: Is the language of the test items at an appropriate level for the test takers?

3. Item content and response behaviors: Are the contents of the items and the behaviors required for answering, appropriate for the test takers?

4. Directions to test takers: Are the directions for each subtest clear and complete? Are there separate directions for each group of items that need to be introduced separately? Are sample (i.e, practice) items given where needed? The difficulty of the directions is relevant here.

5. Testing time: Is the time required for testing appropriate for pupils like yours? If there are time limits, are they appropriate?

6. Layout, print, and illustrations: Do the visible characteristics of the materials make it easy for the pupil to do the test? Are the print and illustrations clear and large enough? Is the content of the illustrations familiar to the pupils? Is the material for each test item adequately separated from the material for other items?

7. Item difficulty: Covered in the module on rating the curricular relevance of a test.

E. Procedural features

Sources of information on the feature: Publishers' catalogs, examiners manuals, directions to pupils, directions for scoring, other sample materials

Relevant test uses: All, except where noted

1. Qualifications of test administrator: Are the qualifications of the test administrator made clear, and are your personnel qualified?

2. Directions to the tester: Are the directions to the tester clear, complete, and easy to use?

3. Ease, speed, and flexibility of scoring: Are the desired options for scoring available? For classroom use of scores, either hand scoring by means of a template (or other objective and efficient key) or machine scoring with very rapid turnaround is needed. For program planning, accountability, or program evaluation, slower scoring may be adequate.

4. Objectivity of scoring: Are the guidelines for scoring so clear that different scorers (a) know what to do and (b) get the same results with the same pupil responses?

5. Curriculum indexing: Does the test publisher offer an optional index relating the specific skills on the test to specific learning activities or to the lessons and exercises in several series of appropriate instructional materials? This is relevant when test scores are to be used to support instructional decisions in the classroom.

6. Availability of alternate test forms: Are two or more parallel forms available? This is important when pupils are to be tested more than once, as in pre- and post-testing.

F. Reporting and interpreting test scores

Sources of information on the feature: Technical manuals, examiners' manuals

Relevant test uses: All, except as indicated

1. Choice of score reports: Are the levels of score reports that you want available? Scores are needed for individual pupils on individual objectives for diagnosis, prescription, and ongoing progress verification. Classroom scores for single objectives are useful to teachers for instructional planning. Class, grade, building, and program level scores may be useful for accountability, needs assessment, or program evaluation.

2. Choice of score types: Are the types of scores that you want available? Mastery (pass/fail), domain (% correct), and percentile norms are possibilities. Grade level equivalents should not be used. Are the scores given in a form that is easy to use and interpret?

3. Score report or record forms: Are the forms for reporting or recording scores easy to use and appropriate for the test use?

4. Guidelines for decision making: Does the publisher give usable advice on how to make decisions about individual pupils on the basis of a combination of information sources, some of them being test scores?

5. Cautions: Does the publisher give information on the limits to interpreting test scores? Information on the amount and sources of error in measurement is useful, as is information on the types and probability of decision error.

6. Appropriateness of norm groups: Are the pupil populations for norming the test meaningful to compare with your pupils?

The following section, pages 4.7 to 4.19, provide step-by-step instructions for comparing tests on the above kinds of features.

Checklist Step 1:

Select test features to evaluate.

The first step is to decide which characteristics to use for comparing the tests' overall technical and practical merits. You can simplify the task by eliminating two sets of features:

- 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, prescriptive curriculum indexing and availability of alternate forms may be eliminated from the judging process when there is to be a 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 student 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 initial 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; test which do not offer optional prescriptive curriculum indexes may be screened out and the remaining tests later compared on the quality of their curriculum indexes.

A worksheet is provided (pages 4.21 - 4.22) in this module for organizing your comparison of tests. In using this kind of form, the first step is to write in the first column of the sheet the names of the features which you want to use for comparing the practical and technical quality of tests for the given testing situation.

Checklist Step 2:

Rate the importance of the features to be compared, and record the ratings on the worksheet.

A test's suitability to meet your needs depends more heavily on some of its features than on others. Three degrees of importance in features have already been mentioned:

- Exclusionary features - ones that are necessary for a test to have it it to meet your needs. These are used in a pass/fail fashion to exclude clearly unacceptable tests.
- Irrelevant or unimportant features - ones that have just been crossed off the list of characteristics to be evaluated.
- 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, after they meet minimum levels of acceptability, as mentioned under Step 1. Also included are all of the other aspects of tests which make them relatively more or less practical and technically sound. These are the features which have not been crossed out on the worksheet plus any you have added.

Now judge the relative importance of these remaining features and assign importance ratings, or weights, to them. A simple scheme would assign a rating of 2 to more important than average features and a rating of 1 to the ones of average importance. This scheme has the advantage of simplicity, but it may not be sensitive enough to the real (in your judgment) differences in importance of test features. A three level weighting system, like this

3 = most important

2 = average importance

1 = useful, but not so important

will recognize a broader range in the value of tests' characteristics.

The later, overall rating of a test is influenced by the importance

weight of each feature. The point of having both exclusionary features for screening tests at first and "importance" weights for adjusting the influence of features on the overall rating is this: we want to keep the less important features from adding up in the final analysis to overcompensate for the absence of essential and more important ones. This principle - Don't let the minor test features dominate the comparison of tests - should guide the test selection process. As noted above, a feature that is of minor importance for one test use may be essential for a different use.

The different audiences and users of the tests 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 use, and of the effects of testing on pupils' motivation and morale.

Checklist Step 3:

Enter the names of the tests to be compared at the top of the worksheet and then duplicate the form.

In the spaces at the top of the worksheet enter the name, form, and level of each test to be evaluated. For use in filling out the rest of the worksheet, write an abbreviation of each test's name in the column labelled Abbreviated Names.

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.

Checklist Step 4:

Find the evidence, e.g., in the sample materials, 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, any relevant stimulus materials, etc. Not all specimen sets are organized the same way, 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 will be noted in the next step.

Find the evidence for the first test feature in all of the specimen sets.

Checklist Step 5:

Arrange the tests in descending order of merit or quality on the first feature. Enter 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 which other 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 zeros on that characteristic. You will have to decide locally on how little merit a test can have on a feature and still be worth ranking above zero. For example, you may decide that reliabilities below .6 are as bad as having no reliability data at all. Then you would rank all tests with no reliability figures or with figures below .6 as zeros, 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.

Checklist Step 6:

How to handle ties and small diferences in ranking tests on a single feature.

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. Imagine a case like this: for the feature concerning number of items per objective, one test has the most appropriate number of items. Two other tests have an equal, and somewhat less suitable, number of items per objective. A fourth test has a still less suitable number. Any test with an unsuitable number of items per objective would have been eliminated in the prior screening, so there should be no zero rankings for this feature.

For this feature the best test will receive a first place and the least suitable one a fourth place. The tests that are tied in the middle will both be ranked (2 + 3) ÷ 2, or 2.5. On the line of the worksheet for that feature draw a circle that includes the spaces for the second and third places, write the abbreviated names of the two tied tests in it, and write 2-1/2 or 2.5 in the circle. 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 deserve different rankings for test features that are very important, but they do not for features that are less important. A second rule of thumb is that small differences in merit deserve the same ranking for features that are judged subjectively or on which different judges disagree a great deal. For features that have clear, objective evidence, small differences in quality are a firmer basis for assigning different rankings.

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 much easier.

Checklist Step 7:

Repeat Steps 4-6 for all of the other test features to be evaluated.

Compare the tests, one feature at a time, and record their rankings on a feature before going on to evaluate the next one. When problems or questions arise, note them in the right-hand column of the worksheet under "Notes." They can be resolved later by conferring with other test raters. Staff members with special expertise should be assigned special features to evaluate, so one person need not rate all of the features.

Checklist Step 8:

Summarize the rankings of all tests by weighting them and transferring them to the Final Results table on the worksheet.

The next step toward an overall comparison of the tests is to transfer the rankings to the summary table at the upper right of the worksheet. The rankings will be recorded as tallies in the Final Results Table and will be weighted according to the importance of their respective features.

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 for that test 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 first place column of the table. Two tests that are tied for second and third place on that feature (thus 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 not acceptable evidence for that same feature would have three tallies entered in the right hand column of the table. All tallies are written on the line of the table opposite the respective tests' name.

Checklist Step 9:

Check your work before proceeding.

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 it is not, re-do Step 8 on a sheet of scratch paper column by column, instead of feature by feature. Again, verify your work be seeing if the number of tallies is equal and correct.

The product 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. These overall profiles will be compared next as the index of tests' technical and practical quality.

Checklist Step 10:

Compare the profiles of rankings of the tests. Decide whether some have markedly better profiles. Select the better ones for detailed curricular analysis, eliminate any that are markedly worse than the others, and keep the others for possible future reference.

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. To transform the tallies into percentages, simply divide the total number of 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 number of tallies which fall in each box. Adding across for each test, the percentages should sum to 100% (plus or minus rounding error).

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 places. 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 physical measurement. At this stage of test selection the purpose is to screen out tests that have markedly lower quality on the features which your program considers relevant.

If there is no obvious break between the higher ranking and lower ranking tests, you may select and screen on the basis of your resources for carrying out an additional 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.

The methods in this module are meant to help you find, screen, and evaluate tests to suit your special situation. The overall 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.

HOW TO SELECT A TEST

Comparing the technical and practical merits of tests

Checklist

____ 1. Select test features to evaluate, using the CSE list of features as a guide.
____ 2. Rate the importance of the test features to be compared, and record the ratings on the worksheet.
____ 3. Write the names of the tests to be compared at the top of the worksheet, and then make copies of the form for the various test raters.
____ 4. Find, in the sample test materials for all tests, the evidence for the first test feature.
____ 5. Arrange the tests from best to worst on the given feature. Record these rankings in the body of the worksheet.
____ 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.
____ 7. Repeat Steps 4-6 for all other test features to be evaluated.
____ 8. Summarize the rankings of all tests in the Final Results Table at the upper right of the worksheet.
____ 9. Check to make sure that the total number of tallies per test in the Final Results Table is equal.
____10. Compare tests' profiles in the Final Results table. Eliminate tests that are markedly worse. Select the better ones for detailed analysis of their congruence with your local curriculum (see module III).

CSE WORKSHEET FOR COMPARING TESTS' TECHNICAL AND PRACTICAL FEATURES

Month/Year _____

Rater(s) _____

Step 3: Names/Forms/Levels of Tests Being Compared	Abbreviated Names	Steps 8-10: Final Results: 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

4.23

CSE WORKSHEET FOR COMPARING TESTS' TECHNICAL AND PRACTICAL FEATURES

Month/Year 3/79

Rater(s) M. Choy (except #4-rated by evaluator)

Step 3: Names/Forms/Levels of Tests Being Compared	Abbreviated Names
Test A (primary level)	A
Test B	B
Test C	C

Steps 8-10: Final Results: Total of Weighted Rankings for Each Test

1st	1-2	2nd	2-3	3rd	3-4	4th	4-5	5th	Not Acceptable -Zero-
7	2	3							2
6	2	6							6
		6							13

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					Zero	NOTES
		Best	Second	Third	Fourth	Fifth		
1. Clarity of skills	3	A	B				C	The objectives make Test A easier to teach toward. The judgments are more clear cut for Test B. Also, Test B uses the same criteria for all pupil responses.
2. Objectivity of scoring	3	B	A				C	
3. Guidelines for decision making	3	A	C				B	If the scoring of Test C were more convincing then its decision rules would be, too.

4. 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.
5. ~~Surface-fairness~~	lined out to indicate that this feature was used earlier in the test selection process to exclude unacceptable tests				
6. Pilot testing	3	A	C	B	
7. Curriculum indexing	2	B		A, C	
8. Alternate forms	2	A/B 1.5		C	(1+2)÷ 2 = 1.5

4:25

CSE WORKSHEET FOR COMPARING TESTS' TECHNICAL AND PRACTICAL FEATURES

Month/Year _____

Rater(s) _____

Step 3: Names/Forms/Levels of Tests Being Compared	Abbreviated Names	Steps 8-10: Final Results: 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

163

TEST A

Description of test: Pupils are given pictures and/or specific oral directions on how to respond. The oral directions, which are very simple, are designed to try to guide pupils' responses so that they show specific grammatical or conceptual skills. Responses are taped for later scoring. The test is based on behavioral objectives, and each item is scored only for the skill it is supposed to test. The test has several levels of difficulty roughly corresponding to certain grades

---

1. Clarity of skills being measured

The following three objectives are a typical sample of how the Teacher's Manual for Test A describes the skills on the various levels of the test.

-Given a picture of a familiar object (or color or shape, etc.) the pupil will say its name.

-Given a picture of persons or familiar objects along with a question word as a prompt (from the set who, where, what, why, how, and when), the pupil will make up a question about the picture that exhibits the correct word order and verb forms.

-Given a picture of two or more familiar objects, the pupil will describe their similarities and differences.

-Given a picture of familiar objects and the oral prompt, "Where is the (name of object)?" the pupil will respond with a phrase or sentence that uses the appropriate preposition of location (e.g., in, near).

2. Objectivity of scoring

Items for each objective are scored only for the specific detail or concept they are designed to measure. Directions for scoring each skill give examples of correct answers, incorrect answers, and problematic answers along with explanations of each. For example, for the items that ask the pupils to describe similarities and differences of objects, the Manual provides acceptable and unacceptable answers for the scorer to use as models and scores the responses on the basis of meaning (i.e., conceptual correctness), not surface grammar.

3. Guidelines for decision making

Two levels of decision making are described by the publisher for using Test A: placement decisions about individuals and diagnostic decisions about individuals' specific skills. The publisher suggests that pupils who get total scores below 20% on the test be classified as failing, 20-75% as partial mastery, and over 75% as having mastered the content.

At the level of the individual objective, scores of 75% or less (i.e., 6 or less correct out of the 8 items per objective), indicate that the pupil needs more practice on that skill, the amount of practice depending, of course, on how many errors the pupils made. These decision rules show that a pupil classified "partial mastery" still needs work on specific oral skills.

For pupils with scores near a borderline between two levels of classification, the publisher advises using other information for interpreting the test scores. She suggests the following procedure: for pupils whose test

scores are near a borderline of classification, give them the benefit of the doubt if their teacher's judgment of their overall proficiency supports the higher level of classification. Whenever teacher and tester's judgment clearly conflict, retest, and if the conflict continues, work it out with the teacher.

4. Reliability of scoring

The published version of Test A was administered to 500 pupils each in grades 4 and 8 who were representative of (your students') age, SES, geography, and dialect range in the United States. Pupils were given both of the equivalent forms of the test. Their responses to the first form were scored independently by two judges, and expressed as failing, partial mastery or full mastery. Results for the two scorers were compared for two conditions: using the test responses alone and using test responses plus classroom teachers' judgments. For the first condition the judges agreed in 83% of the cases. In 84% of the cases, judges' classifications of pupils' responses to one of the test forms agreed with their own classifications of the same individuals' responses to the other test form, showing how equivalent the two test forms are.

5. Surface fairness

A review of the items during test development for surface fairness is not reported for Test A. The test buyer must study the test item by item to judge whether the pictures, instructions, and scoring favor one group over another or are offensive to any group.

6. Pilot testing

The field test version of Test A contained about twice as many items as were needed for the production version. This draft of the test was piloted on a nationally representative sample of students of the targeted age, SES, geography, and dialect range. Five thousand of these pupils, 1,000 at each of grades 2, 4, 6, 8, and 10 took part in the study. Faulty items were identified by test administrators, test scorers, and by the developers, who examined pupils' responses. These items were either repaired or discarded, and the production form of the test was created from the resulting pool of items.

7. Curriculum indexing

None

8. Alternate forms

Two parallel forms of each level of Test A are sold. Purchase of the second form is optional.

Test B

Description of the test: The pupil is shown a selection of pictures and asked to chose one to tell a story about. The pupil tells the story, which is taped for later scoring, and the procedure is repeated eleven times more. The pupil's ten best speech samples are selected for scoring, and scoring is carried out according to a point system which gives a specific number of points for each word, phrase, etc.

---

1. Clarity of skills being measured

Since the items are open-ended and unstructured, they measure (or elicit) all of the speech skills at once. To improve pupils' performance on the test, the Teacher's Guide describes exercises for practicing such things as intonation patterns, noun inflection, verb inflection, basic vocabulary, contractions, asking questions, describing events, giving oral directions, and many more.

2. Objectivity of scoring

Scoring is done by recording pupils' oral responses, transcribing them later, and assigning points according to a standard system. A specified number of points is given for each word, phrase, clause, modifier, partial sentence, and sentence. The manual contains step-by-step procedural directions for deriving each pupil's measure of oral fluency, as well as ten pages of examples of speech samples and how to score them. The examples are chosen to illustrate the basic units (words, phrases, etc.), as well as cases where the scoring might not be obvious, such as fragmentary utterances,

code switching, and non-standard dialect.

3. Guidelines for decision-making

Although individual record forms are provided on which several types of composite score can be recorded, guidelines are not provided for making classification or prescriptive decisions about individuals on the basis of those scores.

4. Reliability of scoring

The production versions of both forms of Test B were given to a sample of 1,000 pupils that was nationally representative of locale, dialect, and age. Five-hundred of the pupils took the same test form on two occasions, about a week apart. The reliability coefficient for test-retest was .90. The other pupils were tested on both test forms during one day. Agreement of individuals' scores on the two forms was sufficient to earn a reliability coefficient of .88.

Data are not yet available on the reliability of Test B with other groups, but the test manual gives step-by-step directions for doing reliability studies locally and for norming tests locally.

5. Surface fairness

Items for Test B were composed by testing specialists who had at least two years of experience in classroom teaching at the respective grade levels. All levels of the test were then reviewed by a national panel of language teachers and representatives of the target student groups. Reviewers eliminated items for which a test taker would need special geographical, ethnic, or socio-economic experiences, or which seemed to portray any social

groups in a stereotyped or prejudiced light. The remaining items were put together into test forms.

6. Pilot testing

Pilot testing of the items for Test B is not reported or described.

7. Curricular indexing

Although items and scores for Test B are not cross-referenced to published English and language arts series, the Teacher's Manual contains ten pages of directions for preparing and conducting prescriptive activities which are designed to improve pupils' performance on the test. Some of the target skills have been mentioned above under #1, Clarity of skills.

8. Alternate forms

Alternate forms, consisting of different sets of stimulus pictures are optionally offered for purchase.

TEST C

Description of the test: The pupil is shown a set of four pictures which depict a sequence of events and is asked to tell a story which mentions the action in each picture. Pupils' responses are tape recorded for later scoring. Scoring is done by making an overall judgment of the quality of pupils' utterances.

---

1. Clarity of skills being measured.

Test C calls for conversational speech samples which are then scored in one overall judgment. Specific component skills are not identified.

2. Objectivity of scoring.

The scorer is told to listen to the pupil's entire story and categorize the speech as follows:

Failing: the meaning and sense of pupil's utterances in English is usually unclear. Although some words may be identifiable in English, they are put together in a fashion that obscures the intended meaning.

Partial mastery: the meaning is clear from as little as sometimes to as much as often, but the grammar shows moderate to severe lapses which impede communication. Pupil's speech is hesitant, halting, or labored.

Full mastery: The meaning is generally clear, and the lapses in grammar do not often block communication. Pupil's speech is generally smooth, not halting.

3. Guidelines for decision making

The pupil is ready to start learning to read in English when all of the following requirements are met:

- The pupils' score on Test C is above the failing level;
- The pupil has the skills that are taught in any standard reading readiness program;
- The pupil is reading words and comprehending text in the primary language at a level that is appropriate for her grade or age.

4. Reliability of scoring

The reliability of Test C, and the effectiveness of the Holistic Scoring Method, is strongly supported by a tryout of the test forms. When two independent judges scored the tapes of a sample of 100 test takers at each test level, their classifications of the pupils agreed 70% of the time. The level of agreement did not vary by more than 5% across levels of pupil.

5. Surface fairness

The items and scoring rules were reviewed by a panel of Hispanic, Oriental, and American Indian educators from California, New Mexico, Arizona, and Colorado. Reviewers were asked to study the materials for possible regional, cultural, or racial biases. Bad items were either revised or eliminated.

6. Pilot testing

The Manual, with its directions for test administration, scoring, and interpretation were piloted in at least three programs for each of

these language groups in California: Spanish (Mexican-American), Chinese (Mandarin), and American Indian. After testing at least 25 pupils in their own programs, educators in these programs suggested revisions of the Manual to improve its clarity and practicality.

7. Curriculum indexing

None

8. Alternate forms

None

Glossary

- An achievement test measures what the student has learned (as in an academic field -- chemistry, English -- or as in basic skills).

- An aptitude test measures a student's potential in a specific area, e.g., academic, scientific, clerical, language, etc.

- Behavioral objectives are goals of instruction which are stated in terms of precisely identified, observable behaviors. The behaviors are indicators for teachers or evaluators of pupils' learning.

- Content or Curricular validity establishes the correspondence between the content of a test and the content of the course for which the test is used.

- Construct validity concerns the psychological qualities a test measures, i.e., the relation between a test and explanatory concepts or theoretical constructs. By both logical and empirical methods, the theory underlying the test is validated. For example, if students with a high level of responsibleness are found to be more willing to acknowledge their mistakes or inappropriate conduct, such behavior may be explained by the construct "responsibleness."

- A control group, in educational experiments, is a group to which the test group is compared. The control group mirrors the test group as closely as possible in every aspect except that the control group does not receive the "treatment" given to the test group. Differences in the group's scores can then be attributed to the differences in treatment.

- Correlation is a commonly used measure of relationship between two variables or paired facts. It shows the extent of similarity in direction and degree of variations in corresponding pairs of scores on two variables. It ranges in value from -1.00 for perfect negative relationship through 0.00 for none or pure chance to +1.00 for perfect positive relationship.

- A criterion, in the classical psychometric sense of the word, refers to a level or standard of performance. However, the word criterion takes on a somewhat different meaning in the expression criterion-referenced test where it signifies a class of behaviors.

- A criterion group is a group of individuals who possess the skills or attributes that one is attempting to measure.

- A criterion-referenced test determines the extent to which a student has mastered a specified domain of behavior (criterion behavior).

- Criterion-related validity may be established by examining how closely students' performance on a predictor test parallels their performance on a criterion measure such as grade-point average, proficiency ratings, or another test.
  - Concurrent validity is based on the relationship between a predictor test and a criterion measure when both variables are assessed in essentially the same time period.

. Predictive validity is obtained through dual measures separated by a span of time. If the predictor test successfully foretells students' performance on the criterion measure, it is predictively valid.

- Cultural bias in measurement refers to factors in item development, test administration, and interpretation of results which favor or penalize members of specific cultural groups.

- Curricular relevance exists when measurement coincides with the school's goals. A useful distinction is made between curricular and instructional relevance. Instructional relevance exists when tests measure what is actually taught in the courses. Thus, instructionally relevant measurement may be more sensitive to actual practice than is measurement of the curriculum on paper.

- Decision errors in measurement generally refer to two particular areas-false positives and false negatives.
  . A false positive error occurs when one incorrectly advances an examinee (say, to a master status), believing the examinee to possess certain skills or characteristics when, in fact, the person does not.
  . A false negative error occurs when an examinee is not judged to possess a skill or characteristic which, in fact, the examinee does possess.

- Descriptive validity refers to the extent to which a test accurately describes the attributes which it claims to measure. Content validity is one aspect of descriptive validity, but non-content variables in the affective and psychomotor domains are included as well.

- Discriminating power refers to the ability of a test to differentiate among individuals who possess different levels of skill in the attribute being measured.

- Domain-referenced tests assess examinee performances with respect to a well circumscribed area (domain) of learner behaviors and subject-matters to which a set of test items are referenced.

- Domain selection validity addresses the relevance of the behavioral domain that has been chosen for a criterion-referenced test. This procedure is similar to the construct validity associated with norm-referenced tests.

- Domain specifications establish the limits of learner behaviors and subject-matter content being measured by a domain-referenced (criterion-referenced) test.

- Equivalent forms reliability is established by giving two forms (equivalent or parallel) of a test to the same person and determining the consistency or agreement of the results.

- Face validity refers to one's instinctive appraisal of a test; does it look valid. Although not based on any sound analysis, if a test does not look valid, it loses credibility in the viewer's eyes.

- A frequency distribution is an arrangement of scores gathered from a group of individuals to show the number of scores (frequency) falling within various intervals (distribution) on the measurement scale being used.

- High internal consistency (reliability) occurs when most items on a test measure essentially the same thing. It consists of high correlation of scores on the different items within the test.

- Item analysis refers to any one of a number of processes used in test construction to determine the effectiveness, difficulty, and discriminating power of an item.

- Item difficulty is determined by the percentage of individuals who get an item right. If ninety percent of the examinees were to answer an item correctly the item would be easy. Conversely, if only ten percent of the examinees were able to answer it correctly, the item would be difficult.

- Kuder-Richardson formulas constitute a widely used method of establishing the internal consistency (reliability) of a test based on item inter-correlations (KR-20) or estimates of such intercorrelations (KR-21).

- A local norm, as opposed to a national norm, is based on the performance of examinees in a particular, generally not widespread, area.

- A mean is the average score received on a test. It is calculated by adding all the scores and dividing that sum by the number of scores.

- A minimum competency test measures a particular minimum number of skills (competencies) necessary to function effectively in the task area which the test measures.

- A norm is the average (expected) score on a test for the members of a particular group.

- A norm group is a group of previously examined individuals from which one establishes a norm.

- Normal distribution refers to an ideal frequency distribution in which the scores cluster around the mean, then taper off at the extremes. The phenomenon is represented by a bell-shaped curve.

- A norm-referenced test compares a student with other students. Norms for locally (classroom), regionally, or nationally sampled comparison groups may be established to interpret how one student compares to other students.

- An objective test is characterized by having a list of correct answers which allows the scorer to avoid subjective evaluation of the student's performance. Multiple-choice, true-false, and matching item tests are examples of objective tests. Short answer and completion item tests may involve some subjectivity, hence are typically not considered objective.

- An operational definition is an indicator or measure of a concept such as achievement or self-esteem. The indicator must be in terms of some measurement or observation, like a test score.

- Process measures are best understood when compared to product measures. Process measures deal with variables which occur during instructional activities, such as the methods the teacher uses, the teacher's educational background, etc. Product measures deal with variables which constitute the outcomes of instruction, i.e., that which the student has learned.

- A proficiency test measures students' expertise in a certain area, such as in basic skills, e.g., reading, mathematics, or in a vocational specialty, e.g., dentistry.

- A random sample is a sample drawn without bias from a specified population usually on the basis of a table of random numbers.

- A raw score is the first quantitative result obtained when scoring a test (frequently the number of correct responses).

- Reliability refers to the consistency of a test's results.

- Sensitivity to instruction, in criterion-referenced measurement, indicates the extent to which student performance on an item can be improved by (effective) instruction.

- The standard deviation (S.D.) is a measure of the dispersion or variability of scores around their mean.

- A standard score is a score expressed as a deviation from the mean in terms of the standard deviation of the distribution (raw score minus the mean, divided by the standard deviation).

- A standardized test is one which has (1) a set of prescribed directions for the test's administration, (2) definite rules for scoring, and (3) norms for score interpretation.

- Stanines (STAndard NINE) are a unit of a standard score scale which divides a distribution of test scores into nine segments. The mean is five and the standard deviation is two.

- Statistical significance describes an event's chance probability. When, for instance, two means are different and the difference is greater than that which would be caused by chance alone, the difference is said to be statistically significant.

- Test-retest reliability occurs when the same test is readministered to the same students after a time interval and produces consistent results.

- Test specifications describe the behavioral domain being assessed by a criterion- or a norm-referenced test. These specifications typically range in descriptive precision.

- Validity refers to the extent to which a test accomplishes the task for which it was intended.

MAKING, CHOOSING AND USING TESTS:

A PRACTICUM ON DOMAIN-REFERENCED TESTING

FACILITATOR'S GUIDE

Submitted to

National Institute of Education

James Burry, Linda G. Polin
and Eva L. Baker

OB-NIE-78-0213
P4

Eva L. Baker
Principal Investigator

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

This project was supported in whole or in part by the National Institute of Education, Department of Health, Education and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

For Limited Use. August, 1980.

GENERAL INTRODUCTION TO TRAINING SESSION

1. Description of materials in the participant notebook:

   a. Modules 1 and 2; dealing with test development
   b. Modules 3 and 4; dealing with test selection
   c. Glossary of terms used in testing and instruction.

2. Purpose of training session:

   a. To introduce participants, through discussion and activities, to test development procedures. This session consists of activities in domain-referenced test specification and construction. It will involve the materials in (1) Module 1 - Domain-Referenced Testing, in which the domain specifications are laid out. Participants develop a domain specification during the session. (This kind of specification generates rules for the construction of test items specifically geared to the instructional domain); and (2) Module 2 - The Item Rating Scale, which provides for systematic examination of a pool of items to see how well they fit the intentions of the domain specification that was developed to guide their generation. Participants will rate a sample of items for their goodness of fit with their specifications.

   b. Modules 3 and 4, dealing with test selection, are intended to provide guidance to people who want to select (standaridized) tests. Module 3 provides step-by-step provedures for comparing tests' relevance to a given curriculum, as well as practice exercises in the comparison process. Module 4 provides step-by-step procedures for comparing tests' technical and practical merits, as well as practice exercises in the comparison process.

3. Preamble to Modules 1 and 2, Domain-Referenced Testing.

   Domain-referenced testing can help in the development of tests which satisfy several important criteria:

   a. Publicness - all involved understand what material will be covered in instruction, what is expected, what will be required and studied, what will be tested and how, how the test results can be used.

   b. Economy - The tests are economical in terms of money, student and teacher time, student and teacher anxiety.

   c. Instructional Sensitivity - the tests are responsive to instructional intervention; they have a specific decision-making purpose.

d. Meaningfulness - the tests are of significance and value to those who give them as well as those who take them.

e. Emphasis - given their deliberate congruence with instructional concerns, the tests are intended to supplement, and not to supplant, other less formal measures used by classroom teachers, such as judgment and observation.

Other advantages and purposes of domain-referenced testing which are related to the above attributes:

a. Stating expectations to students (and the process of teachers and other staff writing the domain specifications):

- reduces anxiety on the part of the student
- takes the mystery out of the test and the testing procedure
- allows both teacher and student to concentrate on teaching and learning
- places or affirms the responsibility of students to direct their own learning to a defined body of knowledge

b. Keeps the teacher focused on his or her own most important concepts or facts and allows the teacher to concentrate on the most effective processes which will bring about learning (because the content has already been decided upon).

c. Adaptability - the content remains the same, while the processes of presenting the materials can vary depending on the age, skills, interests, and background of the students (and of the teacher).

d. Fairness and content validity to students and to the teacher; the process helps ensure validity of tests; teachers and students know what is expected of them; tests provide a more valid measure of what students know.

e. Teachers have control of what will be taught (e.g., content must be teacheable to the particular class of students).

f. The test supplements teacher judgment and provides an excellent context for feedback to the student (what the student has learned as well as what still needs to be learned) as well as to the teacher.

g. The process of specification, item writing, and item validation helps foster a sense of local ownership of the testing process, in that the validity and reliability of the tests are influenced by the local curriculum and the context in which it operates.

MODULE 1: DOMAIN SPECIFICATIONS. Approximate time - 1 1/2 hours

A. INTRODUCTION

Step 1:

a. Introduce yourself (and any other staff working with you)

b. Find out the range of people in the group; e.g., how many teachers? principals? administrators? researchers? evaluators? testing specialists? university personnel? etc. The mix of people in the group should determine the tone and focus of the presentation.

Step 2:

a. Describe the scope of the module and the activities involved.

b. Purpose/Focus:

- what is a domain?
- why is it specified?
- how is it specified (point out that there will be a group exercise in which a domain specification will actually be written)

Step 3:

a. Examples of domain-referenced test content. These examples are intended to give the group a general idea of how domain specification focuses on those aspects of the domain in which students will be taught and tested.

1. Writing a paragraph (or identifying a paragraph)

   a. main idea
   b. supporting details
   c. sentence form; first sentence indented
   d. spelling, punctuation, grammar, handwriting (for constructed response)

2. Writing a friendly letter (or identifying one)

   a. format (date, salutation, closing, signature)
   b. content (friendly, newsy, inquiry, thank-you)
   c. spelling, etc. (if the letter is a constructed response)

3. Addition facts

   a. recognition of whole numbers (and concepts of whole numbers)

b. single-digit whole numbers
c. word problems (to what extent will language be involved in the teaching examples, in the particular problems students will solve, and in the test problems)

4. Identification of mammals

a. self-regulating body temperature
b. usually with body hair
c. nursing of their young
d. many species, including humans

5. Identification of triangles

a. three sides
b. straight sides
c. closed shape

Step 4:

a. Amplify the triangle example to suggest instructional implications of domain specifications; i.e., discuss domain specifications as a means of identifying critical features of instruction and testing. In this example, you will be discussing some of the critical features that could be used in teaching and testing students in identifying triangles. For example:

What are the critical features that define triangles?

- three sides
- straight sides
- closed shape

What kinds of discriminations do you want students to be able to make?

- isosceles triangles
- equilateral triangles
- right triangles

What conditions or barriers do you wish students to be able to cope with?

- triangles standing on their bases as opposed to their vertices
- use of color as a distractor

b. Encourage the participants to respond to the above kinds of questions relating to triangle identification so that the whole group evolves the critical features in the identification.

Keep the group focused on the notion of what you describe as testable must also be teachable.

If there is a blackboard in the room write up the group-developed identification features; or use a blank transparency.

Step 5:

Explain and discuss the components of domain specifications as they are presented in the participant materials:

- point out the materials in the prose section describing domain specifications, pp. 1.1 - 1.8. These materials may be read later.
- point out the activity worksheets the participants will be using in the group exercise, pp. 1.9 -1.19.
- point out the sample domain specifications, pp. 1.20 - 1.33; the annotated cognitive domain taxonomy, p. 1.34; and the suggested sources of objectives and goals, pp. 1.35 - 1.37.

Step 6:

Acknowledge that domain specifications or criterion-referenced test specifications can come in a variety of forms and still contain the same information. The participants may be familiar with Popham's Instructional Objectives Exchange (IOX) abbreviated specifications, for example, or they may have district versions. In discussing the components (e.g., domain description, content limits, etc.), you can refer back to the components generated for the triangle example.

Step 7:

Discuss the components of domain specification. For this step, use the overhead transparencies numbered 1.1 - 1.12 as an aid in your presentation. These transparencies correspond to (1) pp. 1.4 - 1.8 in the participant materials and describe:

- the domain description (transparency 1.1)
- content limits (transparencies 1.2 and 1.3)

- distractor domain (transparency 1.4)
- response criteria (transparencies 1.5 and 1.6)
- format (transparency 1.7)
- directions (transparency 1.8)
- sample item (transparency 1.9)

and (2) pp. 1.20 - 1.21 and 1.26 - 1.33 in the participant materials and show fully worked examples of domain specifications for:

- grade 5 English, selected response (transparency 1.10)
- grade secondary Life Science, selected and constructed responses (transparencies 1.11 - 1.12)

These transparencies are intended to give the group an idea of the features of domain specification, what their task will be in the exercise when they write a domain specification, as well as what their written domain specification could look like.

B. GROUP EXERCISE - WRITING A DOMAIN SPECIFICATION

Step 8:

Now that the group is ready to write a domain specification, have them make the decisions described on pp. 1.9 - 1.10 in their materials; these decisions relate to:

- the subject area and objective which will provide the content of the specifications. The objective chosen should be one of the four examples listed on p. 1.11.
- the grade level of the intended testing audience.
- the difficulty level of the test. See the taxonomy on p. 1.34 in the participant materials.
- the kind of item to be treated in the specifications. The group should decide if they want to develop specifications for a selected response or a constructed response.

You should keep the group working together as these decisions are made. For purposes of the exercise, it will probably be easier for you and the group if a selected response decision is made. If the group cannot agree on a grade, suggest some central grade, such as 6 or 7, as a compromise.

Step 9:

When the group decisions have been made, have the participants turn to the worksheets in their materials. These worksheets begin on p. 1.12 and run to p. 1.19. These are the pages on which the participant write their specifications.

The group should be evolving the specification together; i.e., consensus should be reached in each area before going on to the next. Obviously, things may bog down unless you keep the group moving along. It will help if you use a blackboard to develop the specification while the group work on their notebooks; or use a blank transparency.

While the groups are working on a particular component of their specification, e.g., domain description, content limits, etc., re-show the corresponding transparency that goes with this module, so that the participants have a quick reference to the features of each task in the specification process.

Step 10:

Each disagreement among group members should be an opportunity to stress the instructional and testing linkages in domain specification; e.g., Do you want to teach this? Can you teach this? Is that actually a different objective?

Step 11:

When the group reaches the content limits section briefly reiterate the difference between selected and constructed response item types, and the differences involved in writing content limits for each (different emphases). Based on the previously made group decisions, the participants will be working on only one of these item type.

Step 12:

When the group reaches the distractor domain sections, the difference between selected and constructed response item and specification needs should be briefly reiterated. Again, based on the earlier decisions, the group will be working with only one of these components.

Step 13:

When the group reaches the format, directions and sample item sections, do not let them pass these off too lightly. Consider the complexity and structural needs of the directions. Consider the labelling and constructing of the item set-up. Consider the importance of these sections in constructed items (e.g., oral tests, performance tests, demonstrations).

If there is time left at the end of the module, you may wish to take the group back over the materials and this time work in the item mode that was not used the first time through. This might also be an incentive for the group to keep moving along during the first go through.

Step 14:

Wrap-up and questions and answer period. If there is time remaining, you might start a discussion of how this module might be used in subsequent in-service.

MODULE 2: ITEM RATING SCALE. Approximate time = 1 1/2 hours

A. INTRODUCTION

Step 1: (only if you did not run Module 1 with the group)

a. Introduce yourself (and any other staff working with you)

b. Find out the range of people in the group; e.g., how many teachers? principals? administrators? researchers? evaluators? testing specialists? university personnel? etc. The mix of people in the group should determine the tone and focus of the presentation.

Step 2:

a. Describe the scope of the module and the activities involved.

b. Purposes/Focus:

- the implications of domain-reference testing; to include specifications, item development, item review for fit with the written specifications.
- how may one judge the quality of the items written based on the domain specification?
- what kinds of judgments can be made about items and how can they (or the specifications, if appropriate) be revised?
- need for a judgment system that suggests areas of revision in the item or the specification, as opposed to judgments by experts or review panels which may not fulfill the needs for systematic and informative review. The Item Review Scale (IRS) is intended to help fill this need. (Point out that there will be a group exercise in which items will be compared with their domain specifications and rated for goodness of fit.)

Step 3:

Explain and discuss the components of the IRS as they are presented in the participant materials:

- point out the materials in the prose section discussing the IRS, pp. 2.1 - 2.4. These materials may be read at a later time.
- point out the explanations of the IRS on pp. 2.5 - 2.11. These pages will be discussed during the module.
- point out the exercise activities on p. 2.12 (rating sheet to be used in the exercise); and pp. 2.13 - 2.17, or 2.21 - 2.25, or

2.26 - 2.31 (domain specifications and items to be rated). Additional copies of the rating sheets are on pp. 2.18 - 2.20.

Step 4:

a. Discuss the components of the IRS. For this step, use the overhead transparencies numbered 2.1 - 2.13 as an aid in your presentation. These transparencies correspond to pp. 2.5 - 2.11 in the participant materials. Each transparency breaks out a single component of the item rating scale as follows:

- suggested rating guidelines for using the IRS (transparency 2.1)
- domain description (transparency 2.2)
- content limits - selected response items (transparency 2.3)
- content limits - constructed response items (transparency 2.4)
- distractor limits (selected response only) (transparency 2.5)
- response criteria (contructed response only) (transparency 2.6)
- format (transparency 2.7)
- directions (transparency 2.8)
- sample item (transparency 2.9)
- linguistic complexity (transparency 2.10)
- thinking complexity (transparency 2.11)
- overall item rating sheet (transparency 2.12)
- guide for interpreting ratings (transparency 2.13)

b. While you are showing the transparencies, point out that the statements listed under each category are intended to guide the raters in his/her consideration of the degree of match between the test writer's intention (i.e., the specifications) and the item itself.

For participants who are not familiar with the selected or constructed response item forms, a brief explanation of the difference in items and domain specifications will be necessary (refer to Module 1).

c. Point out to the participants that the IRS uses each component of domain specification as a rating section; i.e., domain description, content limits, distractor limits or response criteria; format; directions; and sample item. In this way, an item can be rated against the specific rule prescribing each feature. Point out that in addition to the above domain specifications, which are identical to those treated in module 1, the IRS also includes two other categories for rating the match between an item and the specification. These are linguistic complexity and thinking complexity. These categories are intended to provide a screen for those features of an item that are not described and limited in the domain specification, but which are likely to have an effect on the difficulty of the item. In such cases, the intentions of the specification may not be realized. Point out that there is also an overall item rating sheet, and a guide for completing the rating.

d. When you have shown and briefly discussed each of the transparencies once, have the participants turn to the overall item rating sheet in their materials (p. 2.10) and re-exhibit the corresponding transparency (2.12). Explain the weightings and the ten-point scale used to rate each section in the IRS. At this point, you may also wish to re-exhibit the transparency showing the suggested rating guidelines (2.1), and use both transparencies as an aid in discussing the scoring system.

   Then have the participants turn to the guide for interpreting ratings (p. 2.11) while you re-exhibit the corresponding transparency. Use the guide for discussing the decision rules for item revision (or domian specification revision if appropriate).

B. GROUP EXERCISE - RATING ITEMS AGAINST THEIR DOMAIN SPECIFICATIONS

Note: This activity can become bogged down in discussion and argumentation as raters express their scores for any given category. This can be (and should be at first) used to refer again to the value of domain specifications in limiting and defining the content and conditions of tests (and by implication, of instruction). You might suggest that differences voiced may be the result of differences in personal value or emphasis placed on the same factor by people in the group. You might also point out that when all raters are working in the same content and value system (e.g., at the level of the individual school or possibly the district) the IRS has yielded high levels of reliability or consistency among raters.

Keep redirecting the participants to the domain specification that they are rating items against in the exercise. Often participants rate against their own standards rather than those given them to use for the exercise.

Step 5:

Have the group decide on the sample domain specification and items that they would like to rate. Depending on the particular subject matter familiarity of the majority of the group, you may wish to use the English-punctuation specification (pp. 2.13 - 2.14) and its corresponding items (pp. 2.15 - 2.17); or the Elementary mathematics-set theory specification (pp. 2.21 - 2.22) and its corresponding items (pp. 2.23 - 2.25); or the Elementary science-geology specification (pp. 2.26 - 2.27) and its corresponding items (pp. 2.28 - 2.32).

Step 6:

After the group has decided on which specification/items to rate, have them remove the worksheet from their materials (p. 2.12). Briefly go over how to use the rating sheet in rating the item against the specification.

If there are no questions, you can begin to walk through the exercise.

Step 7:

Have the participants read the specification they have selected to work with and the items that go with the specification.

Step 8:

When the participants are familiar with the specification/items, have them focus specifically on the first item wirtten for the specification.

Go back to the transparencies that go with this module and flash the first category (i.e., domain description - transparency 2.2), and lead a discussion of the item through the statements listed for that category.

NOTE: In the past, some participants have mistakenly begun to rate the sample item that is included with the specification. Avoid this problem by ensuring that all participants are looking at the first test item written for the specifications, and not the sample item that was used to help generate the pool of items.

Step 9:

Ask different people for their rating of the item against the category of domain description. If many participants are hesitant or if there is too great a range in responses, show the suggested rating guidelines on the overhead again, and ask participants to elaborate the flaws and the degree of seriousness with which they think these flaw violate the domain specifications.

Step 10:

Eventually, ask the group if anyone is unhappy or unable to live with a particular score (choose the most popular rating). Bring the group to consensus and have then enter the rating on their sheets.

Still using the first test item, move on to the next rating category exhibiting the appropriate transparency, repeat step 9, and eventually bring the group to consensus on the second category.

Repeat the above process for each category on the rating scale until the first item has undergone the complete cycle. Then compute the overall rating for that item.

NOTE: Make sure that the group includes the weighting system where appropriate (i.e., content limits; distractor domain/response criteria; and thinking complexity).

Sometimes a participant may uncover a flaw in the item-specification match that either reflects mainly upon the specification or that is best dealt with in a different category of the IRS (e.g., thinking complexity). For subjects best dealt with in other categories, ask the participant to wait and see if the other category picks up that particular concern.

Step 11:

Point out that the interpretation guidelines include suggestions to reconceptualize or restructure both the item and the specifications; that is, it is not always the item that must be revised. The item may be appropriate for the intended testing situation and the specifications are somewhat off target and therefore need to be clarified.

- Don't be appalled at low ratings that may come up. Raters often tend to be harsh critics; also, the example items were deliberately marred in certain places.

- For example, if your group used the English-punctuation specification and items to rate (or perhaps even if they did not) point out the potential problems with items 1 - 3 for this specification in terms of thinking complexity. The first prompt for items 1 - 3 is often identified as being too complex in terms of required comprehension for the sixth grade level. Participants often bring up the difficulty of recognizing the loudspeaker as a voice requiring quotation marks.

Step 12:

Ask the participants to turn to the interpretation guide (p. 2.11) and discuss the rating given to the item rated in steps 9 and 10 and how this rating might be interpreted. Discuss the changes in the item or the specification that the group may recommend.

If there is sufficient time, have the participants rate another item or two from the specifications previously used in steps 9 and 10.

If there is time to do this, point out the importance of rating each item separately against all the rating categories, rather than rating an item on the first category, then rating another item on the first category, and so forth. The principle reason for rating each item separately against each category before beginning to rate another item is that we want to have each item rated independently of the other items in the pool. When items are rated against the same category at the same time, the ratings given to one of the items may tend to influence the ratings given to another item. In this way potentially good items may be kicked out.

If participants do have time to go through the second rating process, save some time to discuss the various ratings given.

Step 13:

Wrap up and question and answer period. Remind participants of the extra rating sheets and additional example materials that they can use later.

MODULE 3: COMPARING THE RELEVANCE OF TESTS TO A GIVEN CURRICULUM

Approximate time for this Module: 2 hours if participants rate the entire sample test, 1 1/2 hours if the leader only walks them through the first two objectives.

A. INTRODUCTION

Step 1:

Intoduce yourself

Step 2:

Discuss the scope and sequence of this module, which will consist of:

1. short oral introduction and discussion
2. introduction to materials
3. demonstration of procedures with the first two objectives
4. participants finish rating the sample test
5. discuss how to use the output from the ratings

Step 3:

Intended audiences: teachers, curriculum specialists, and anyone who will coordinate or supervise test selection.

Step 4:

Set the context for this module; for example:

- the functions and limits of testing in evaluation
- mention earlier stages in test selection, such as:
  - deciding whether to test
  - picking domains to test
  - choosing test functions (survey, diagnosis, mastery, aptitude, etc.)
  - identifying available tests
  - screening tests from information in secondary sources
  - ordering specimen sets
  - screening sample test materials for disqualifying characteristics such as cultural, geographical, or dialectal biases
  - comparing the surviving tests' technical and practical merits
  - the importance of attending closely to tests' content when choosing a measure to use in a local program

Step 5:

Describe contents of notebook

A written explanation of the checklist for review, self instruction, and teaching to others. Explain that you will be giving this information orally in the workshop, do not have them read it (pp. 3.1 - 3.25).

Checklist to compare tests' relevance to a given curriculum (p. 3.26).

A partially filled-in worksheet to use in the workshop exercise (pp. 3.27 - 3.30).

A fully worked example worksheet, based on the exercise materials. It serves also as feedback for the exercise (pp. 3.31 - 3.34).

A blank worksheet for local reproduction or modification (pp. 3.35 - 3.37).

All-American Test of Reading Comprehension, a mock-up (pp. 3.38 - 3.46).

Step 6:

EXERCISE: Using these materials: the All-American Test, the partially filled-in worksheet, and the one-page checklist. (90 minutes)

Step 7:

Using a transparency of the worksheet, skim through the steps in the checklist. Introduce the final ratings as what you ar working toward.

Step 8:

Briefly describe Step 1. Discuss the importance of Step 1, which for this exercise has been carried out already. The listing of an imaginary curriculum has been done for the participants. Step 2 also has been done for them. At this point emphasize the importance of teachers and curriculum coordinators as test raters - that is, people who are intimately familiar with the real curriculum and with pupils' capabilities.

Step 9:

Have them carry out Step 3, counting and recording program skills.

Step 10:

Lead the group in doing importance ratings, Step 4, on the first two of the program skills on the partially worked rating sheet.

Step 11:

After taking questions, have them finish rating the importance of the sample program skills.

Step 12:

Mention Steps 5 and 6, duplicating the rating sheet, labelling it, and indexing the test items, which have already been done on the partially worked sample worksheet. Have them do Step 7, counting and recording the number of items on the test.

Step 13:

Walk them through Steps 8 and 9 with the items for the first two objectives. Items for the first objective are designed to illustrate the dimensions of content, and the two sets of items for the second objective illustrate item format and solution processes.

Step 14:

Walk them through Step 10, multiplying the ratings, for the first two objectives. At this point acknowledge the detailed nature of the process and discuss these two points:

- the process is easier to do than to hear about, and it becomes much easier with a little practice
- the process is designed so that all of the many component judgments are recorded and their effects carried through to the final ratings of a test. In more impressionistic or intuitive methods of judging tests, the component decisions may get ignored, lost, or mis-remembered.

Step 15:

Have the group carry out Steps 8, 9 and 10 on the rest of the sample test materials. Try to meander through the audience to answer questions, get a feel for the process, etc.

Step 16:

At some predetermined time, stop everyone for questions and discussion. Then walk them through computing the final ratings.

Step 17:

On a transparency, go over the test ratings (i.e., Grand Averages, Indices of Coverages, and Indices of Relevance) and discuss how to use them for comparing tests' curricular relevance (i.e., what their indices mean for purposes of comparison).

CLOSING OR TRANSITIONAL REMARKS

MODULE 4: COMPARING THE TECHNICAL AND PRACTICAL MERITS OF TESTS

Approximate time for this module: 90 minutes maximum

A. INTRODUCTION

Step 1:

Introduce yourself

Step 2:

Discuss the scope and sequence of this module, which will consist of:

1. short oral introduction and discussion
2. introduction to materials
3. demonstration of procedures with a couple of test features
4. participants finish rating the sample materials
5. discuss how to use the Final Results Table

Step 3:

Intended audiences: Anyone who will take part in or supervise the process of selecting tests, including:

- teachers
- test administrators
- testing specialists
- evaluators

Step 4:

Set the context for module, for example:

1. the role of testing in evaluation
2. mention earlier stages in test selection, such as:
   - deciding whether to test
   - picking domains to test
   - choosing test functions (survey, diagnosis, mastery, aptitude, ets.)
   - identifying available tests
   - screening tests from information in secondary sources
   - ordering specimen sets
   - screening sample test materials for disqualifying characteristics such as cultural, geographical, or dialectal biases.

Step 5:

Describe contents:

1. A written explanation of the checklist for review, self-instruction, and teaching to others. Explain that you will be giving this information orally in the workshop; do not have them read it (pp. 4.1 - 4.19).

2. Checklist to compare the technical and practical merits of tests (4.20).

3. Three copies of a worksheet for carrying out the steps on the checklist.

   - Two blank worksheets, one for reproducing or modifying locally (pp. 4.25 - 4.26) and one for use in the workshop exercise (pp. 4.21 - 4.22).

   - A fully worked example that is based on the exercise materials. It serves as a model and a feedback to the participants (pp. 4.23 - 4.24).

   - Descriptions of three imaginary tests, Tests A, B, and C (pp. 4.27 - 4.36).

Step 6:

EXERCISE: Using these materials: the checklist, the partially filled-in worksheet, and the information on fictional Tests A, B, and C.

Step 7:

Give a general idea of the procedure. Encourage discussion throughout.

- First, orally refer to the familiar practice of ranking things as best, second, third.

- Then, introduce the notion of test features.

- Combine these two notions into ranking tests with respect to individual features.

Step 8:

Using a transparency of the worksheet, introduce the worksheet, and then show how it serves to record your rankings of tests, feature by feature.

Step 9:

Discuss Checklist Step 1, noting that the annotated list of features (in the participant's materials) is an aid to this step, and noting that the choice of features to be evaluated is made locally, varying from one testing purpose to another.

Introduce the features which are pre-selected for the exercise, noting that #5 is excluded.

Step 10:

Introduce the importance weighting of features (Step 2) and walk through the weighting of the features #1 (Clarity of skills) and #7 (Curriculum indexing).

Step 11:

Have them do Step 3.

Step 12:

Walk through Steps 4 and 5 for the first feature.

Step 13:

Walk through Steps 4, 5, and 6 for feature #8 (Alternate forms).

Step 14:

Let them compare the tests on the remaining features and record their rankings (Step 7).

Step 15:

After they finish, explain how to summarize the rankings in the Final Results Table (Step 8).

Step 16:

Explain how to check the accuracy of the Final Results Table (Step 9).

Step 17:

Discuss how to make decisions with the Final Results Table (Step 10).

CLOSING OR TRANSITIONAL REMARKS