DOCUMENT RESUME

ED 189 168 TM 800 341

AUTHOR Hambleton, Ronald K.; Simon, Robert A.
TITLE Steps for Constructing Criterion-Referenced Tests. Laboratory of Psychometric and Evaluative Research Report No. 104.
INSTITUTION Massachusetts Univ., Amherst. School of Education.
PUB DATE Apr 80
NOTE 64p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Cutting Scores; *Guidelines; Scoring; *Test Construction; *Test Format; Testing Problems
IDENTIFIERS Domain Specification; *Test Content; Test Manuals

ABSTRACT

The subject of constructing criterion-referenced tests is often researched, but many technical problems remain to be satisfactorily resolved. Foremost, criterion-referenced test developers need a comprehensive set of steps for construction. In this paper, 14 logical steps for building criterion-referenced tests that refer to several different applications and allow for objective and non-objective formats are offered: 1) preliminary considerations; 2) identification of possible content; 3) preparation of domain specifications; 4) review of domain specifications; 5) additional test planning; 6) preparation of test content; 7) preparation of scoring method; 8) test materials review; 9) compilation of final form of test; 10) determination of standards; 11) preparation of report forms; 12) preparation of technical manual; 13) publication of test; and 14) collection of technical data. Four significant contributions of the steps are: 1) use of a priori methods for validation; 2) allowance for use of objective/non-objective test formats; 3) flexibility of steps for use in distinct situations (classroom; district/state; state and national); and 4) comprehensiveness of steps. In addition to the steps, a discussion of rationale for inclusion of each step and guidelines for implementation are provided. (GSK)

Steps for Constructing Criterion-Referenced Tests[1,2]

Ronald K. Hambleton and Robert A. Simon
University of Massachusetts, Amherst

Glaser (1963) and Popham and Husek (1969) were the first researchers to make the case for criterion-referenced tests. Popham and Husek also offered a set of methods and procedures for constructing criterion-referenced tests and interpreting test scores. Since the pioneering work of Popham and Husek in 1969, there have been hundreds of research papers written about technical matters associated with building criterion-referenced tests. For example, the psychometric literature abounds with papers which consider such topics as (1) writing objectives, (2) preparing and validating test items, (3) determining test lengths, (4) selecting test items, (5) assessing the reliability and validity of test scores and decisions, and (6) evaluating tests. Berk (1980), Hambleton, Swaminathan, Algina, and Coulson (1978), and Popham (1978) offer reviews of many of these contributions.

Of course many technical problems remain to be satisfactorily resolved. For one, criterion-referenced test developers need a comprehensive set of steps for building criterion-referenced tests. The availability of a set of steps would increase the likelihood that test developers would consider all of the proper steps and carry them out in the correct sequence. Unfortunately, current models for criterion-referenced test development have several shortcomings. One shortcoming

[1]Laboratory of Psychometric and Evaluative Research Report No. 104. Amherst, MA: School of Education, University of Massachusetts, 1980.

[2]A paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

is that they emphasize the building of tests which use multiple-choice, true-false, or matching questions (Hambleton & Eignor, 1979a, 1979b; Millman, 1974; Popham, 1978). A common criticism of criterion-referenced tests (or basic skills tests, competency tests, or minimum competency tests, as they are sometimes called) is that there is almost a total reliance on objective formats and therefore the tests are limited in the skills they can measure. Many important skills such as writing and speaking can be measured better (and sometimes only) through the use of essays, observational methods, and simulations, to name just three non-objective item formats.

Reliance on objective test items is due to the (relative) ease with which they can be written and administered, to the convenient way in which they can be scored, and to the lack of experience among test developers in using formats for test data collections such as observations, simulations, and work-samples. But, criterion-referenced tests need not consist solely of objective test items. For example, National Assessment of Educational Progress uses a variety of item types in order to provide useful information about the quality of American schools. If criterion-referenced testing programs are to achieve their full potential, more use must be made on non-objective formats so that skills such as writing checks, utilizing the resources of a library, and preparing a resume can be assessed.

Another shortcoming of available models for test development is that they are often specific to particular applications. It would be highly desirable to have a list of steps which is broad enough to guide the preparation (1) of tests at the classroom level (for diagnosis and monitoring student progress), (2) of tests at the district and state level (for program evaluation and remediation) and (3) of tests at the state and national level for use in certification and licensure.

It seems clear then that there is a definite need for a comprehensive set of steps for building criterion-referenced tests. Also, it seems unnecessarily restrictive to offer a set of steps which are limited to a particular format or to a particular application. In this paper a set of logical steps for building criterion-referenced tests that apply to several common (but different) applications and allow for both objective and non-objective formats will be offered. The steps represent a combination and extension of prior work by Tinkleman (1971), Osborne (1973), McKeegan (undated), Sanders and Sachse (1975) and Hambleton and Eignor (1979a). Four significant contributions of the steps are:

1. The use of a priori methods to validate the test blueprint.
2. The allowance for the use of both objective and non-objective test formats by placing the format decision in its proper position in the sequence.
3. The flexibility of the steps for use in three relatively distinct situations, i.e., classroom tests, large scale assessment and occupational/professional licensure and certification examinations.
4. The comprehensiveness of the steps in that they cover the entire process of test development and validation pertaining to the assessment of both knowledge and skills.

Constructing Criterion-Referenced Tests

In this section of the paper a set of 14 steps will be introduced along with a brief discussion of each step. The 14 step model is presented in Figure 1. In most instances the outline is sufficiently descriptive so elaboration in the text is minimized. The text consists primarily of points which need elaboration and additional comments concerning some aspects of the steps.

1. Preliminary Considerations in Preparing a Test

The first step is essential to keep the process focused in a useful direction. A committee which represents those groups which have some responsibility for the test should be formed to oversee the test development process. The committee should address itself to matters such as:

1. the purpose(s) of the test
2. the group(s) to be assessed
3. identification of recipients of test score information and how they will use the information
4. the content areas (specified in general terms) which will be covered by the test
5. the test length specified in terms of the approximate time available to administer the test
6. the amount of time, money, expertise, and personnel available to carry out the test development process
7. a timeline for test development, and assign people and resources to assure completion of each step.

Figure 1. Steps for constructing criterion-referenced tests.

1. Preliminary considerations in preparing a test

   a. State the purpose(s) of the test
      i. Classroom (for example, diagnosis, description, or instructional decision-making)
      ii. Large Scale Assessment (for example, program evaluation, or student remediation)
      iii. Certification and Licensure (for example, awarding of high school diplomas, or controlling entry into occupations and professions)

   b. Identify the group(s) to be assessed and the groups who will receive test score information

   c. Specify the content area to be covered and the approximate test administration time (or test length)

   d. Specify the amount of time, money, and expertise available to complete the test development project

   e. Prepare a list of activities, attach deadlines, and assign people and resources

2. Identification of possible content for inclusion in a test

   a. Form a committee of individuals to carry out the required work

   b. Prepare a first draft of the content (a listing of specific behaviors or topics is desired)
      i. Classroom
         •build from the present curriculum and what is currently taught
      ii. Large Scale Assessment
         •review curricula and textbooks
         •involve individuals with an interest in the scope and direction of the test (for example, parents, community leaders, legislators, school board members, curriculum specialists, principals, teachers, and students)
      iii. Certification and Licensure
         •prepare an initial list of jobs and associated responsibilities and functions (and possibly specify activities, knowledge, and skills at this time as well)
         •complete the list of jobs, responsibilities, etc., with the aid of textbooks, interviews with trainers and practitioners
         •if high school graduation exams, generate content consistent with the purpose(s) of the test

c. Specify the content in "descriptive" objectives (i.e., with sufficient specificity for people to understand the content)

d. Select the most appropriate objectives for additional consideration
   i. Classroom
      - relevant groups (probably teachers but possibly parents and administrators too) can meet to discuss the merits of different objectives in relation to the purpose(s) of the test
      - consensus decision-making, the Delphi technique, and questionnaires are three possible ways of collecting data
   ii. Large Scale Assessment.
      - decision-makers meet to select the content
      - surveys of interested individuals (for example, parents, teachers, principals, and students) can be carried out and the results are used by the committee in making decisions about content
      - a combination of the two methods can be initiated
   iii. Certification and Licensure
      - survey job holders and ask them to rate job components in terms of their importance and frequency of occurrence
      - if high school graduation exams, decision-makers can make a selection of content with the aid of survey data (respondents can be asked to "rank" competencies, and indicate their level of importance)

e. Validate the selection of content
   i. Classroom
      - seek opinions of the test content from teachers, parents, principals, etc. (if suggested revisions are substantial, revise the content and repeat this step)
   ii. Large Scale Assessment
      - seek opinions of the test content from teachers, parents, principals, and community leaders
   iii. Certification and Licensure
      - determine the match (or degree of overlap) between the job specification and the content
      - if high school graduation exams, seek opinions of the test content from relevant decision-makers associations, etc.
   iv. Make necessary revisions and/or additions to the content

3. Preparation of "domain specifications"

a. Organize the validated objectives in a useful way (for example, they can be organized around broad content categories), and prepare domain specifications (or some other type of device for clarifying the scope of content and format to assess performance on the objectives)

b. Determine which objectives can be combined by giving special attention to:
   i. test format (objective vs. non-objective)
   ii. test environment (actual or simulation)
   iii. personnel requirements
   iv. methods of scoring
   v. materials needed and performance aids

4. Review of domain specifications

   a. Identify reviewers and train them in their task

   b. Assess the clarity, completeness (on the validated objectives from step 2 being measured), choice of item format, etc. of the domain specifications.

   c. Revise the domain specifications based on data from 4(b)

5. Additional test planning

   a. Assess the feasibility of including all of the domain specifications in the test (consider the costs and time)

   b. If some must be eliminated, consider the ranking data collected at step 2. Also, consider combining several of the less important validated objectives into one.

   c. With multiple domain specifications, there may be advantages, if simulations are to be involved, to connect them to one another via a common theme or situation.

   d. State the "number of test items" to measure each domain specification

   e. Determine the number of test item writers needed and plan for having them complete their work.

6. Preparation of the "test content" (Do "a" or "b")

   a. Non-objective format
      i. collect performance aids/obtain resources required by the domain specification
      ii. give instructions to item writers along with a copy of the domain specification
      iii. prepare test content, student and administrator directions, aids, props, handouts, and set time limits (if necessary)

   b. Objective format
      i. give instructions to item writers and indicate the number of items to be written
      ii. prepare a draft set of test items and edit them
      iii. prepare a draft set of directions for administrators and examinees

7. Preparation of a scoring method (Do "a" or "b" again)

   a. Non-objective format
      i. choose a scoring method from possibilities specified in each domain specification
      ii. prepare scoring forms (usually both objective and non-objective forms) for process, products, or both
      iii. prepare detailed methods for using the scoring forms and training scorers

b. Objective format
   i. develop scoring keys to reflect item formats
   ii. prepare methods for scoring items

8. Test materials review

a. Content specialists review test directions, content, and scoring; study items for racial, ethnic, and sex bias; and provide suggestions for revision

b. Measurement specialists review the technical soundness of test methods (item quality, validity of scoring, layout, time limits, etc.) and provide suggestions for revision

c. Make necessary revisions based on 8(a) and (b)

d. Try out the test materials on a sample of examinees similar in characteristics to the groups for whom the test is intended

e. Make revisions based on 8(d) and assess test score reliability

f. If revisions are extensive, repeat step 8(d)

9. Compilation of the final form (or forms) of the test

a. Finalize the test directions

b. Compile the final draft of test content (prepare parallel-forms if necessary)

c. Finalize and state the scoring method

d. Provide for test security (this step is not always necessary)

e. Have representatives of minority groups study the items for bias

f. Design and carry out an equating study (from one form to another)

g. Prepare a practice test for administration prior to the test

10. Determination of standards

a. Form a standard-setting committee

b. Select a standard-setting method, train the committee in its use and implement it

c. Assess the reliability of the derived standards across members of the committee or across "parallel" committees

d. Design and conduct a study to address the validity of decisions resulting from the use of the standards

11. Preparation of report forms

a. Prepare an informative reporting form to contain all relevant information and which is written in a style which will be meaningful to those for whom the report is intended

b. Form a committee to review the material from 11(a) and to make necessary revisions and extensions

c. Finalize the report forms

12. Preparation of a technical manual

a. Administer the test to appropriate samples of examinees

b. Assess the reliability of descriptions and decisions of all reported scores. With the judgmental scoring formats it is also necessary to check the inter-rater and inter-observer reliability of both the objective-type and subjective-type scoring criteria

c. Assess the construct validity of descriptions and decisions of all reported scores.

d. Compile norms-tables (if desired)

e. Reassess the cut-off scores, related results (percent masters and non-masters), and their implications and make modifications

13. Publication of the test

a. Finalize item layout and format

b. Print the test, technical manual, along with report forms and an interpretation guide

c. Allow for different cut-off scores in the reporting of results

14. Collection of technical data (over time)

a. Plan to collect item statistics and test score reliability, validity, and norms information periodically

The results of this step should be written up and used as a guide by those who will actually construct the test.

2. Identification of Possible Content for Inclusion in a Test

The outcome of this step is a curriculum or job relevant test blueprint. The precision of the blueprint should be tempered by the importance attached to the test scores. If a test is to be used to make important decisions such as certifying pilots or doctors, or granting high school diplomas, meticulous care should be taken in determining test content. Carefully chosen individuals or groups who have an interest in the test, who may be influenced by them, or who have conent expertise should be represented in the process. If a test is to be used to monitor classroom progress, then somewhat less effort should be expended here unless the curriculum will be put in place across a large number of schools.

First, a committee should be formed to carry out the required work. For classroom tests this committee might include the teacher, but also perhaps other teachers and/or parents as well. For large scale assessment, individuals with an interest in the test should be involved. This might include teachers, parents, administrators, community leaders, etc. For certification or licensure tests the committee would include representatives from professional organizations and the government.

The next task is to prepare an extensive list of possible content. This list can be quite long—even hundreds of objectives. Brainstorming is a good technique for generating a list because no evaluation of the desirability of including any particular knowledge or skill is to take place at this stage. After brainstorming (or during it) the list should be extended. For classroom tests the list can be built

from the present curriculum. Lists for large scale assessment projects should be drawn from available curricula and textbooks but ideas should also be solicited from all those who may have an interest in the test, i.e., parents, citizens, educators, school board(s), the business community, union members, scholars and even students should be surveyed for additional test content ideas. For occupational/licensure tests an exhaustive job list should be drawn from textbooks, curricula, trainers (teachers), practitioners, observational studies, and job analysis studies.

The elements which have been identified for possible inclusion in the test should be put into "descriptive objective" form. A descriptive objective is used so that other people have a clearer picture of what is on the list, i.e., what the objectives mean. A descriptive objective has two components: (1) the behavior of interest, and (2) a partial list of the component skills of the behavior of interest. Two examples of descriptive objectives are given below:

1. Descriptive Objective — Utilize the resources of a library

   Component Skills

   - Use a card catalogue
   - Use a reader's guide
   - Use the reference section

2. Descriptive Objective — Maintain family finances

   Component Skills

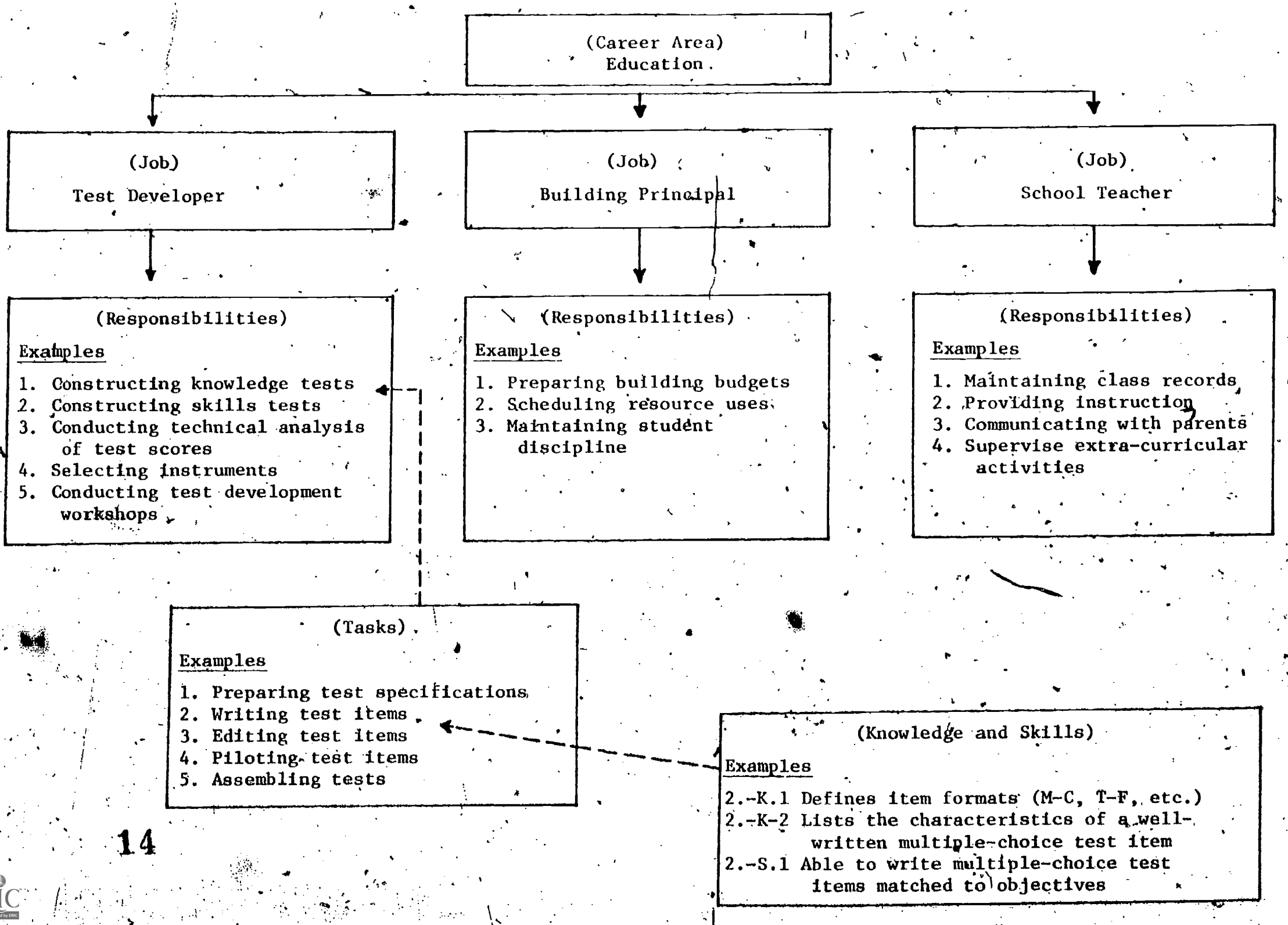
   - Balance checking account
   - Create a realistic budget

Even better, although it may be too time consuming if the testing project is a small one, is the preparation of an "occupational analysis" (i.e.,

the specification of responsibilities, tasks, and corresponding knowledge and skills which define an occupation). An example for the occupation "test developer" is presented in Figure 2.

After the possible content has been extensively listed, the next step is to select the content which is appropriate for inclusion in the test blueprint. If the test is for use in a single classroom, the teacher may be the sole decision maker but other teachers may help out. Depending upon the importance attached to the test parents and/or students might be of assistance as well. If the test is for an entire grade then all interested teachers should be involved in the process. At a meeting to discuss the test blueprint, decisions may be reached via some form of consensus (or close to it) or a group process such as the Delphi technique. A questionnaire could be used particularly if parents are involved, but if the number of participants is small this procedure may be unwarranted. If the test is for a large scale assessment project then a survey of the school and community should be undertaken. The community could be defined as broadly as the test is important; suffice it to say that interested citizenry and those people on whom the test has an effect should be included in the process. The survey should involve a questionnaire which should be a listing of the entire list of descriptive objectives. The respondents should be asked to determine the criticality of each behavior on some form of relative importance

Figure 2. Example of an Occupational Analysis



scale.. When a test is for use in certifying occupational personnel or licensing professionals a similar procedure (i.e., the survey questionnaire listing objectives which asks respondents to judge their relative importance) to that of large scale assessment should be used. In this case the respondents should primarily be practitioners, but an astute test developer may also want to include trainers and consumers in the survey population. A less desirable procedure (and, in fact, a less acceptable method in terms of judicial scrutiny) is one where trainers meet to discuss the merits of one objective over another.

The final step in developing a test blueprint is to validate the selection of the content. At the classroom level the teacher (or teachers) may want to have other colleagues, parents and/or administrators inspect the tentative blueprint and make suggestions for improvement or give their "stamp of approval" to the test outline. If there are very many of these suggestions for improvement the blueprint should go back to those who made it to begin with. If, when developing a classroom level test, the content "validators" are the same as the content "determiners" (a procedure which is not a particularly good one) then it is suggested that the determining and validating procedures be done at least a few days apart from one another. For large scale assessment or certification and licensure tests it might appear that the use of an extensive survey to determine test content automatically produces a valid test blueprint. This is not the case. The results (including relative ranking) of the survey should be compiled into logical (or meaningful) categories and reviewed. Large scale assessment projects should seek

opinions concerning the comprehensiveness, representativeness and relevance of the tentatively selected objectives from teachers, parents, administrators, scholars and community leaders. Tentative blueprints which are to be used in tests for high school graduation should be examined by representatives associated with those groups in society which are effected by the test. Tentative blueprints for certification or licensure tests should be reviewed by knowledgeable teachers and practitioners. Also, careful attention should be paid to existing job descriptions to assure that there is a reasonable line up between the test blueprint and the occupation. Necessary revisions or additions to the test blueprint should be made based on the results of the final reviews of the blueprint. In all cases, the committee which is in charge of the testing project should monitor (and, likely; be involved) in all phases of developing the test blueprint.

3. Preparation of "Domain Specifications"

The outcome of this step is a set of domain specifications (see Popham, 1978). The procedure is exhaustive with respect to each validated objective. It is important to note that no question like, "Is it feasible to test this?" or "Is this domain specification necessary?" should be asked until step five. This step requires expansion of the descriptive objectives into domain specifications.

Each validated objective must be included in at least one domain specification. Validated objectives may appear in more than one domain specification. This may occur as domain specifications become broader and consequently can include more material. Also, a validated objective may have both a knowledge component (which lends itself to paper-and-

pencil maasurement) and a skill component (which lends itself to performance based measurement). In addition to the standards applied in the writing of domain specifications (for methods and examples, see Popham [1978] and Hambleton & Eignor [1979a]) there are some other elements which need to be considered. Domain specifications should be written for both objective (paper and pencil) and non-objective (performance based) items. If the domain specification is for performance based testing then the environment for testing, personnel requirements, possible scoring techniques, and materials and performance aids which are needed for the test should be considered and included in the specification. Two examples are offered in Appendix A.[1] The first is for performance in a "closed domain," i.e., the examinee has relatively limited parameters for acceptable performance. Other examples of closed performance are "filling out an income tax form," "filling out a job application," "making a hospital bed" or "replacing a carburetor." The second one is for performance in an "open domain," i.e., the examinee has a relative freedom in choosing a method of acceptable performance. Domain specifications in this area are more difficult to score but these difficulties are manageable. Other examples of open performance are "leading a group," "handling office work flow," "bedside manner," "writing a newspaper article," etc. It is possible to construct domain specifications in this area and these important areas of human endeavor need not be ignored.

Appendix B is a short introduction to the types of non-objective test formats. This should prove to be an interesting section to those who are interested in going beyond standard paper and pencil objective test formats.

[1]We note here however that the scoring sections of the domain specifications require more work.

4. Review of Domain Specifications

The product of this step is a set of domain specifications of acceptable quality. The domain specifications which were constructed in Step 3 are reviewed for clarity and completeness. Also, the sample test items are reviewed to determine their appropriateness as indicators of the content or behaviors defined by the domain specifications. Finally, the domain specifications are compared to the test blueprint in order to be certain that the validated objectives are adequately covered.

5. Additional Test Planning

The outcome of this step is a reduced set of domain specifications which will be used to prepare the test. Three concerns should be addressed: (1) determine which domain specifications have the most scope within practical limits; (2) determine which domain specifications can be combined into a common thread (or scenario) in order to integrate the test and increase fidelity and representativeness; and (3) the number and type of items. As these three points are considered it is important to keep in mind (a) the purpose of the test and resources available for testing derived in Step 1, and (b) the validated list of objectives derived in Step 2.

In order to make these decisions the classroom teacher can decide solely or in conjunction with others who are interested in the use of the test. Large-scale assessment endeavors and occupational/professional testing programs must rely on a group process to make these decisions. The groups should include (again) all interested parties and almost

certainly will include the committee overseeing the test development process. Decisions concerning the number of items to be used in each domain and in the test should be carefully considered in light of the above concerns but also in order to appropriately maximize the validity of decisions arising from the use of the test.

6. Preparation of the "Test Content"

The outcome of this step is a set of test items drawn from the approved domain specifications. This step is split into two branches: (1) non-objective format—for performance-based items designed to tap examinee skill, and (2) objective format—for paper and pencil items designed to tap examinee knowledge. Only the first branch will be considered here; the second is well known.

The first thing to do is to make sure that the resources which are needed for the test situations are available. Next, instructions should be given to item writers. The instructions consist primarily of the domain specifications but when constructing a situation the item writers will have to tend to other details. In addition to writing directions and items for the examiner and examinee, other standardizing aspects should be articulated, e.g., physical conditions, personnel requirements, number of examinees to be tested simultaneously, specify needed equipment (and its condition)—for both examinee and examiner, etc.

Directions for the test administrator probably should:

1. Specify testing materials and recommend they be checked before testing begins.

2. Describe clearly what an administrator should do and say. Occasionally, it is helpful if directions also mention what test administrators should not say and do.
3. Provide an overview of the testing process.
4. Describe ways for the test administrator to introduce the test and put the examinee at ease.
5. Stress the importance of having prior training (or at least practice) in administering the test.

Directions for the examinee probably should:

1. Address the purpose of testing and why an examinee should perform to the best of his/her ability.
2. Explain each step in the testing process.
3. Address time limits.
4. Explain the scoring system.
5. Introduce performance (or job) aids.
6. Explain the test environment and the amount of realism which is expected.

In composing test items, item writers should adhere rather strictly to the domain specifications-at-hand and strive to set up situations that are as real-life-like as possible within the aforementioned constraints of the testing program.

7. Preparation of a Scoring Method

The outcome of this step is a method for scoring the test. Again, we will not address the procedure one should use in scoring objective tests but rather we will focus on the scoring of non-objective tests. Scoring of non-objective tests can take a variety of forms. Some example formats for scoring tests are presented in Appendix C.

At this stage the item writer should choose from the scoring possibilities articulated in the domain specification. Central to this decision should be what scoring scheme will yield the most valid information within the constraints of practicality. When developing a simulation the item writer may suggest the degree of precision required for satisfactory performance (this should not be confused with standard setting which is addressed in Step 10).

8. Test Materials Review

The result of this step is a group of items which are ready to be compiled into the test. For classroom tests this step need not be elaborate but it should be thorough. All test items should be scrutinized to determine that they do in fact measure the domain specifications of interest and that they do not include any technical flaws. For large-scale assessment and occupational/professional examinations, this step should be treated in its entirety. The items which have been written and their attendant scoring procedures should be reviewed by content specialists for content acceptability and scoring appropriateness and by measurement specialists for technical acceptability and scoring appropriateness. Possible forms for reviewing test items and scoring

non-objective test items are presented in Appendix D. Based on the results of these reviews, items should be left intact (if accepted), discarded (if hopeless) or revised (if possible). The revised items should then be subjected to review again.

Next, the items should be subjected to a pilot test. Careful attention should be paid to all aspects of the testing situation. Areas which should be addressed in the pilot are item statistics (see Popham, 1978; Hambleton et al., 1978), clarity of directions, readability of the test items, speededness, item bias, etc. Reviewers should also check to make sure that the non-objective scoring procedures are articulated well and are working properly (i.e., leading to reliable and valid scores). Also, the scoring choice (from Step 7) should be reconsidered. On the basis of the pilot test, items should again be either left intact, discarded or revised.

9. Compilation of the Final Form (or Forms) of the Test

The outcome of this step is the test in its final form. This entails final editing of the test directions, compiling the items into the test and carefully delineating performance aids. In addition, some final decisions have to be made about the ways in which the test will be scored. In the case of objective tests this procedure is usually rather straightforward (although discussions about the relative weighting of true-false and multiple-choice items often produce lively debates). Decisions about non-objective scoring procedures are difficult and important. A committee consisting of both content and measurement specialists should meet to

determine which scoring procedures are most relevant to the task yet are psychometrically sound. These discussions can best take place in light of the pilot test results. Once the decisions are finalized, directions for scoring and the finalized scoring forms can be compiled into the test.

It may be necessary to consider providing for test security. Depending upon the situation in which the test may be used this may or may not be necessary.

If there are parallel forms it will probably be necessary to design and implement an equating study.

10. Determination of Standards

The matter of standard-setting is a difficult one to deal with. It is clear that all standard-setting methods are judgmental and arbitrary. However, as Popham (1978) correctly pointed out, arbitrary standards are not bad or undesirable if by arbitrary it is meant that a clearly developed plan for standard-setting was prepared, critiqued, and implemented. Readers are referred to Hambleton and Eignor (1979a, 1979b) for two reviews of the standard-setting literature and other references are provided there as well.

11. Preparation of Report Forms

The outcome of this step is a reporting system which meets the needs of those with an interest in the test. A representative committee might meet to determine the form and content of the reports, but this is not absolutely necessary. It is possible to elicit the desires of the various groups in separate meetings, interviews or questionnaires. After

an initial draft is made the report form should be reviewed by the committee. It would be most helpful if sample information were provided in the form. After revision the form should be finalized and made ready for publication. It is unlikely that the committee would have to review the revisions.

This step has had a history of neglect. When all is said and done any test is not worth any more than the information derived and conveyed from it. Careful, even meticulous, attention to this step can have big pay-offs in terms of the usefulness of the test. The reader is referred to Mills and Hambleton (1980) for a thorough and informative presentation of how to report test scores.

12. Preparation of a Technical Manual

The well-known APA/AERA/NCME Standards for Educational and Psychological Tests published by the American Psychological Association in 1974 provides a complete set of guidelines for preparing technical manuals. It suffices to say here that a good test manual should fully describe the test development and norming process, test administration directions, and reliability and validity information in relation to each of the possible uses of scores derived from the test.

13. Publication of the Test

The outcome of this step is the finalized version of the test, administrators manual, technical manual, report forms, and performance aids (if appropriate). While this may seem to be a rather straightforward step the interested reader should see Thorndike's (1971) article on this issue.

If the test is for wide-scale use we suggest that the usefulness of various cut-off scores be reported in the final version. This may greatly enhance the usefulness of the test for different locales.

14. Collection of Technical Data (Over Time)

Regardless of the strengths of a testing program in a particular situation at a given point in time, curricula change, and so do expectations for high school graduation, for entry-level into a profession, job characteristics, and the types of people who are in programs, etc. This means that the psychometric properties of tests will not remain static. Periodic reassessment of test score reliability and validity is essential. And, to paraphrase Bob Linn, norms unlike wine do not improve with age and so norms tables must be updated periodically.

Conclusions and Suggestions for Research

In this paper a comprehensive model for building and validating criterion-referenced tests was introduced. The model is not in final form at this time, but we do feel it can be helpful to test developers in sequencing their activities. We feel equally positive about our support for the use of non-objective formats. Considerable research and development work has been done in industry and the military with these formats. Similar work should be done in education. The formats have much to offer in the way of enhancing the validity of test scores and related decisions.

Additional research should take several directions. First, there is considerable need to substantiate the test development and validation model. This might be constructively done by having test developers (1) check the model for completeness and clarity and (2) match it to the way in which they go about their work (or would if they could choose an approach). Gaps and ambiguities in the model can be identified and used as a basis for making model revisions. Second, there is a need to go beyond the model and provide detailed methods and procedures for carrying out each of the fourteen steps. Without methods and procedures there is not an effective way for applying the model. Finally, more examples of domain specifications in many content areas, like the two in Appendix B, are needed.

Hopefully, some of the ideas and material presented in this paper will encourage others to extend and improve upon our work. We hope so because much work remains to be done and the potential for improving the usefulness of criterion-referenced tests is substantial.

References

Berk, R. (Ed.) Criterion-referenced measurement: State of the art. Baltimore, MD: Johns Hopkins Press, 1980.

Fitzpatrick, R., & Morrison, E. J. Performance and product evaluation. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D.C.: American Council on Education, 1971.

Frederiksen, N. Proficiency tests for training evaluation. In R. Glaser (Ed.), Training research and education. New York: Wiley, 1965.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70. (2nd ed.) Amherst, MA: School of Education, University of Massachusetts, 1979. (a)

Hambleton, R. K., & Eignor, D. R. Competency test development, validation, and standard setting. In R. Jaeger & C. Tittle (Eds.), Minimum competency achievement testing. Berkeley, CA: McCutchan Publishing Co., 1979. (b)

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

McKeegan, H. F. Applied performance testing: What is it? Why use it? Portland, OR: Clearinghouse for Applied Performance Testing, Northwest Regional Laboratory, Paper #1, undated.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutchan Publishing Co., 1974.

Mills, C. N., & Hambleton, R. K. Guidelines for reporting criterion-referenced test score information. Laboratory of Psychometric and Evaluative Research Report No. 100. Amherst, MA: School of Education, University of Massachusetts, 1979.

Osborne, W. C. Developing performance tests for training evaluation. Alexandria, VA: Human Resources Research Organization, Hum RRO-PP-3-73, February 1973.

Panitz, A., & Olivo, C. T. Handbook for developing and administering occupational competency testing. Washington, D.C.: U.S. Department of Health, Education and Welfare, Office of Education, National Center for Educational Research and Development, National Occupational Competency Testing Project, Research Project #8-0474, 1971.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1978.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Sanders, J. R., & Sachse, T. P. Problems and potentials of applied performance testing. Proceedings of the National Conference on the Future of Applied Performance Testing. Portland, OR: Northwest Regional Educational Laboratory, 1975.

Thorndike, R. L. Reproducing the test. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D.C.: American Council on Education, 1971.

Tinkelman, S. N. Planning the objective test. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D.C.: American Council on Education, 1971.

Appendix A

Sample Domain Specifications

(1) Writing Checks for Specified Amounts

(2) Utilizing the Resources of a Library

Objective

Student is able to write checks for specified amounts and to record and balance the transactions on check registers.

Level

Senior High School

Sample Directions for Performance

You have a new checking account at a bank. The checks and register have just arrived in the mail. With the checks it is now possible to pay a few bills which require payment. The checking account was opened with a deposit of $525.90. The bills to be paid are:

(1) Bank Plastics, Inc.	$75.40
(2) Martha's Gas Co.	$12.30
(3) Mortimer J. Snerd	$275.00
(4) Undermountain Utilities	$27.53

You should pay these bills by writing checks, and recording and balancing each transaction in the check register. The checks need not be mailed; just give them to the proctor along with the register when you are finished.

You have fifteen minutes to complete the task.

Content/Behavior Domain

1. The examinee will be asked to write at least three and not more than five checks.

2. The beginning balance will be given as an amount between $100.00 and $999.99.
3. The checking account will be "new," i.e., there will be no checks already on the register.
4. The examinee will give the completed checks and the register to the proctor when finished.
5. There is no restriction on the subtraction problems involved, i.e., the examinee will be expected to borrow (as a subtraction procedure), subtract cents and dollars, and keep the decimal point where it belongs.
6. The checks would be written to fictitious companies or individuals.
7. The examinee will not be asked to overdraw on the account.

Performance Aids and Environment

1. The examinee will be given a check register form with no previous entries.
2. The examinee will be given double the amount of blank checks which are needed to pay the bills. (This is in case certain checks must be voided.)
3. A pen is necessary.
4. The checks should be authentic checks.
5. The checks should be seriated (pre-numbered).
6. Check registers which use stubs should not be used.
7. The environment should be a quiet, unhurried one.
8. The workspace should be adequate.
9. Calculators are not allowed.
10. A blank piece of paper is allowed.

Scoring

Objective Criteria

A recommended scoring key for the performance task follows:

(a) Accuracy

The check	Yes	No
1. Correct date.	___	___
2. Name of payee in the proper space.	___	___
3. Numerical amount in the proper space.	___	___
4. Numerical amount is the correct amount.	___	___
5. Numerical amount written correctly in numbers, i.e., 51.27.	___	___
6. Numerical amount written correctly in words.	___	___
7. Signature in proper place.	___	___
8. Proper name signed to check. (Middle name may be deleted or abbreviated.)	___	___
9. Reason for check noted in "memo" section. (optional)	___	___
The register		
10. Transaction entered on register		
a. check number	___	___
b. date	___	___
c. payable to	___	___
d. correct amount	___	___
e. amount in correct column	___	___
f. amount correctly deducted from prior balance	___	___

(B) Time

1. Task completed in allotted time.	___	___
2. If less than allotted time—total elapsed time.	___(in minutes)	

Subjective Criteria

(A) Rating scales (place a "√" in the appropriate column)

	Unacceptable	Acceptable
1. Handwriting is legible.	____	____
2. Numbers are clear.	____	____
3. Signature is executed in a consistent manner.	____	____
4. Register is kept orderly.	____	____
5. Register is legible.	____	____

A student is identified as a "master" of this skill if his/her performance on the Objective Criteria is 100% (excluding #9) and 100% of the subjective ratings are in the "acceptable" category.

Objective

The student is able to use the resources of a library to gather material for preparing reports on selected topics.

Level

Senior High School

Sample Student Directions

You have been assigned the topic of "Whales and Their Struggle for Survival." To complete the assignment you must find source material in the library in order to write about the topic. The details of your task are as follows:

- You have two (2) hours to gather material.
- You have the entire library at your disposal.
- You should select the material you need and check it out according to library procedures. No more than eight (8) items may be checked out.
- Reference books may not be checked out so if you want to get information from them, then you must take notes and bring the notes out of the library.
- You are not allowed to photocopy material.
- You may not ask the librarian questions during the assignment.

You will be observed during this task and may be asked questions by the observer concerning your activities. At the end of the two hours you will be asked to do two things:

- Give your notes and the material you have checked out to the observer,

- Write a brief explanation of why you chose the particular materials that you did.

Content/Behavior Domain

1. The examinee will be assigned a topic that is of general interest and for which there is material available in books, journals, newspapers, and reference books. Examples of topics are "Whales and Their Struggle for Survival," "The Design and Safety Features of Modern Airplanes," "The Career of Henry Aaron," and "History of the Olympics."
2. The examinee must have borrowing privileges so that material can be checked out and evaluated.
3. After checking out the material (at the end of two hours), the observer will ask the examinee to write a brief rationale for the selection of each piece of material. Preparing rationale statements should require an additional ten to twenty minutes.
4. The examinee will be allowed to use the entire library to locate material.
5. The examinee will be told:
   - that note-taking is acceptable,
   - to locate material for use in writing a report on the assigned topic,
   - of the presence of an observer,
   - that questions will be asked concerning their activities.
6. The observer will collect the notes and the material which were checked out at the end of the two hours.

Performance Aids and Environment

1. A library of suitable size is to be used. School libraries with more than 10,000 volumes would normally be acceptable.
2. The library should have substantial information on the selected topic. ("Substantial" means that there is enough material in the library so that someone who possessed the skill could collect enough material to prepare the desired report.)
3. The examinee must have (at least temporary) borrowing privileges. The material which has been checked out may be returned within a half hour after the test in order to allow the next group of examinees access to the same material.
4. The examinee should have a notebook and pencil or pen.
5. The observer should be as unobtrusive as possible but may interrupt for brief periods in order to assess examinee performance.

Scoring (Several possibilities are given)

Objective Criteria

(A) Time (expressed in minutes)

1. Time used (start to finish) _____
   - Amount of time used in locating material. _____
   - Amount of time used taking notes. _____
   - Amount of time off task, e.g., bathroom, talking to friends, etc. _____

(B) Accuracy

1. Locates material from classification numbers in the card catalogue. (number)

- citations _____
- "finds" _____
- N/A (checked out, missing, etc.) _____
- citations checked out _____

2. Goes to correct place of items. (check one)

directly	one error	two errors	> two errors	gives up

(C) Accomplishments (number)

- items checked out of the library _____
- pages of notes taken _____
- citations (or at least classification numbers) written down _____
- items perused (or used) but not checked out _____

(D) Effort

1. Number of steps taken [as measured by a pedometer] _____

Subjective Criteria

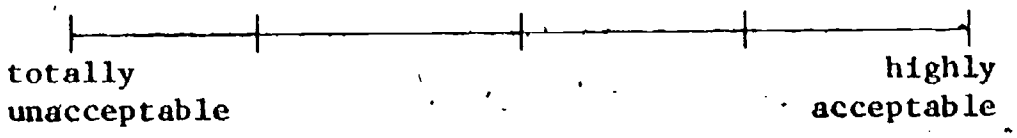
(A) Self-Rating Scales

1. Ease of the task
2. Suitability of selected material

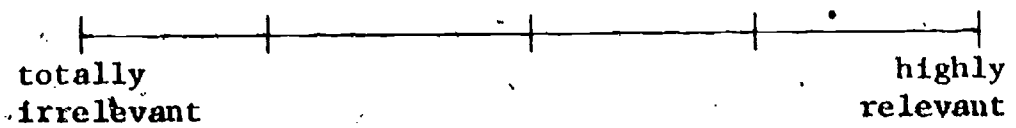
(B) Observer Rating Scales

Rate the candidate (by placing a "√" at the appropriate spot) on each scale.

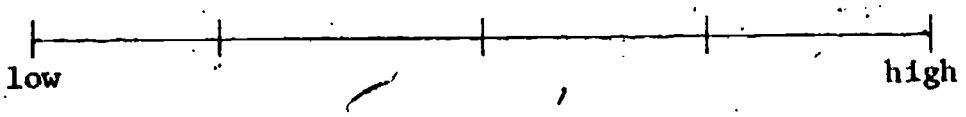
1. Rationale statements for materials.



2. Relevancy of the materials.



3. Diversity of materials.



Appendix B

Types of Non-Objective Tests

Types of Non-Objective Tests

The purpose of this section is to describe several non-objective test formats. We will not attempt an exhaustive categorization process but we will provide a framework, some common terminology and descriptions.

Frederiksen (1965) lists seven methods of obtaining measures for use in assessing examinee performance:

1. Solicit opinions—This can be accomplished formally or informally. Examinees (or individuals who know them) can be asked to provide ratings of performance.

2. Administer attitude scales—When the content of the scale is relevant to the behaviors of interest, the two measures should be (at least) moderately related.

3. Measure knowledge—This can be done via the development of a paper-and-pencil test. It is not usually sound to assume that knowledge of facts and principles is closely related to skill in performing a task.

4. Elicit related behavior—An example of this would be to have a student edit or rewrite writing samples as a test of English composition ability.

5. Elicit "what I would do" behavior—A common problem with this approach is that real-life problems generally don't present themselves in a multiple-choice format, or, at least one which is presented with insufficient information.

6. Elicit lifelike behavior—This involves using a simulation or at least a situation that is set-up by the test developer.

7. Observe real-life behavior—This is impossible to standardize. Often real-life behavior is used as a criterion for examinee success/unsuccess (e.g., supervisor ratings). Caution is warranted due to the fact that many intervening and uncontrollable variables may enter into the situation.

Objective test formats are commonly used to assess knowledge (method 3) whereas non-objective test formats can be used to assess skills (methods 4 and 6).

Panitz and Olivo (1971) and Fitzpatrick and Morrison (1971), among others, supply a scheme for categorizing tests using non-objective formats. Table B.1 provides information for comparing four types of tests: recognition tests, simulation tests, work-sample tests, and project/products tests. What follows is a brief description of each:

1. Recognition Tests—This is sometimes called an "identification test" and measures the examinee's skill in recognizing the essential characteristics of a process or product by naming the object, describing the operation, and/or delineating the function. For example, a telephone repair person could be presented with a picture of a telephone set-up and be asked if the system was set up correctly. A diesel mechanic could be asked to identify the parts of an engine and their function and could even be asked to do it in a pre-specified order. We can include in this category certain problem-solving tests. For example, a licensure test for medicine could present the examinee with a medical history and the results of certain diagnostic tests. The examinee may be asked to interpret the findings and present possible treatment or recommend further testing.

   Identification tests can be given orally, in writing or even by computer. Careful attention should be given to sampling a variety of representative tasks from the test blueprint. The scoring of these tests should be objective and should clearly differentiate mastery/non-mastery proficiency. These tests have the advantage of being resonably easy to construct, administer and score, but do not readily measure Frederiksen's category number six: elicit lifelike behavior.

2. Simulation Tests—In simulation tests an examinee carries out realistic tasks in a setting which simulates a real situation. Role-playing is often an essential ingredient of a simulation. For example, a "psychologist" (examinee) may be asked to treat a "client." A managerial trainee is confronted with an "in-basket" on his/her desk and be asked to respond to a variety of plausible problems. Computer, or other, "games" which present interactive problems to "generals," "economists," "managers," etc., can be grouped within this category of testing. Simulations are often used when the situation is too large (e.g., economics) or amorphous (e.g., management) to lend themselves to be readily measured. An even more compelling use of simulations occurs when the job presents a health or safety hazard. Airline pilot training makes extensive use of simulations as does the training of astronauts. The health professions are increasingly utilizing simulations of clinical conditions. Programs which train people in dangerous professions, e.g., ship's captain, workers who deal with high voltage electricity, etc., frequently utilize simulations.

Table B.1 Types of Tests

Characteristic	Recognition Tests	Simulation Tests	Work-Sample Tests	Project/Product Tests
Useful Situations for Application	1. large groups of examinees 2. economy is important	1. where the situation is too large and amorphous to have "real" situation 2. factors under consideration must be limited 3. where health or safety is a factor	1. when on-the-job observation is possible 2. where the work in question can be accurately observed 3. primarily used with skilled or semi-skilled workers	1. where process is not important 2. where a variety of processes are acceptable 3. when test development and administration costs are limited
Examples	1. identify parts of a diagram 2. point to specified components 3. identify functions of various components	1. role playing 2. games (computer & otherwise) 3. in-basket 4. secretarial tests	1. troubleshoot and repair 2. production output, e.g., machinist, secretary	1. artistic projects 2. sports contests 3. science fairs
Validity for Determining Proficiency	1. low for skills 2. high for knowledge	1. moderate/high for skills 2. moderate/low for knowledge	1. high for skills 2. moderate for knowledge	1. moderate/high for skills 2. moderate for knowledge
Response Modes	1. paper and pencil -multiple choice -fill in blank 2. oral 3. computer interaction	1. paper and pencil 2. computer interaction 3. oral 4. manipulative	varies; depends on actual job requirements	varies; the only response is the product

-B3-

Table B.1 Types of Tests

Characteristic	Recognition Tests	Simulation Tests	Work-Sample Tests	Project/Product Tests
Scoring Modes	1. objective	1. objective, e.g., did/did not do 2. subjective, e.g., observer ratings	1. objective, e.g. output, waste, accuracy, etc. 2. subjective, e.g., rating scales, ranking, etc.	1. objective, e.g., measure tolerances, product works/does not work, amount completed, etc. 2. subjective, e.g., artistic merit
Process/Product Evaluation	does not apply	process and/or product	process and product	product
Costs	relatively inexpensive to develop, administer and score	expensive to develop, administer and score	expensive to develop costs vary to administer (often it is on-the-job time). relatively expensive to score	inexpensive to develop costs vary to administer costs vary to score
Fidelity	low	high	high	moderate/high
Useful as Instructional Device	yes	yes	no	yes
Comments		The test constructor must strive for maximum fidelity within allotted resources		

Simulations often entail a variety of response modes: paper-and-pencil, oral, computer, manipulative, etc. This can present difficult scoring problems. Also, caution is warranted in assuming a degree of relationship between simulated performance and performance with actual equipment and people under realistic conditions. Finally, careful attention should be paid to which tasks are simulated. An effective task analysis may alleviate many difficulties with respect to test validity but a sample of isolated tasks or series of tasks may not be a valid sample of the total job situation.

3. Work Sample Test—While these tests may appear in some ways to be similar to simulations the essential difference is that it requires that the individual demonstrate proficiency by doing a series of tasks or completing a piece of work under actual work conditions. This is the most "realistic" type of test available and has the highest face validity. For practical purposes the test often consists of a sample of a job. For example, it may not be feasible for a T.V. technician to rebuild an entire set so we may observe her/his troubleshooting and repair skills. Work sample tests have primarily been used in the past with semi-skilled or skilled workers. We see little reason, however, for limiting their application.

   It is difficult to standardize this type of test but it is not an impossible undertaking. When the sample of work is an appropriate one these tests can provide reliable and valid estimates of proficiency.

4. Project/Product Tests—This type of examination entails evaluation of only the result of a series of tasks. Something is presented and evaluated. Science fairs, musical or dramatic performances, most athletic competition, art shows, industrial arts projects, etc., are only a few of the types of activities which readily lend themselves to this type of evaluation. Evaluating only the finished product ignores adequately assessing process and examinee knowledge, but nevertheless this type of test is often quite useful and generally very economical.

All four types of tests described above have considerable potential for criterion-referenced test developers.

Appendix C

Example Formats for Scoring Non-Objective Tests

Example Formats for Scoring Non-Objective Tests

The scales which are delineated below are suggestive of the types of scoring formats which are available. Scoring is an important consideration and a difficult responsibility for the test constructor. Depending on the scope of the skill(s) to be evaluated it is unlikely that only one format will adequately measure examinee proficiency. When designing a test, the test constructor should peruse this list to see which scoring procedures can adequately be used to assess proficiency.

The first five types of procedures are relatively more objective than the following five types. Fortunately, there are at least three promising methods for increasing the reliability of assessments:

1. Use several indicators (or measures) of performance,
2. Increase the number of skills to be measured,
3. Thoroughly train (and re-train) observers/scorers.

Objective Measurement

1. Time

This is a measure dealing with the amount of time which an examinee uses in demonstrating a skill.

Example:

Time started _____

Time finished _____

Elapsed time _____

1. Accuracy

These are measures which deal with the correctness of a product or process.

Example 2.1:

Number of typing errors on a ten-minute test: _____

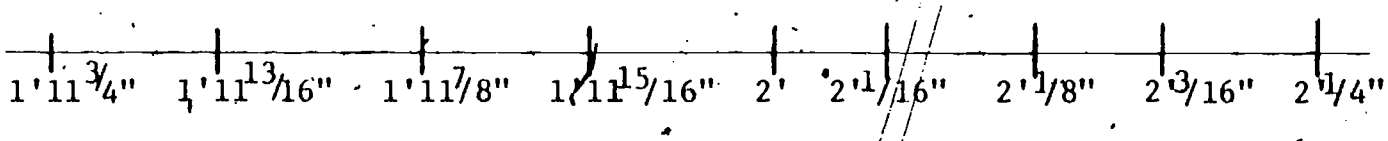
Is the stock cut to desired length (± .01 inch)? _____

Is the blueprint prepared according to the specifications? _____

Example 2.2:

Objective: Wooden bookshelf cut to 1/16" accuracy.

Scale: Wood cut at two feet.



The observer is told to measure the wood and record the dimension. The scoring could be 10 pts. for 1'11 15/16" - 2'1/16", 8 pts. for the next 1/16" from perfect, 6 pts. for the next 1/16" and 0 pts. beyond that.

3. Frequency of Occurrence

These are measures dealing with the frequency of behavior repetition.

Example 3.1:

Within a 20 minute time period, observe the number of times a teacher does the following things (put a check for each occurrence):

	Tallies	Total
Asks recall question	_____	_____
Asks student to read	_____	_____
Provides feedback	_____	_____

Example 3.2:

(For an in-basket simulation). Within a one hour time limit observe the number of times an examinee performs each of the following things (put a check for each occurrence):

	Tallies	Total
Reads something from in-basket	_____	_____
Dictates memorandum to subordinate	_____	_____
Dictates letter to client	_____	_____
Drafts a personal memorandum	_____	_____
Puts information back into in-basket	_____	_____

4. Amount Achieved or Accomplished

These measures deal with the amount of output produced by an examinee.

Example 4.1:

Number of words typed in 5 minutes. _____.

Example 4.2:

Number of telephone inquiries handled in one hour. _____.

Number of times supervisor helps with an inquiry. _____.

Example 4.3:

Wickets packaged in a 15 minute time period:

(Directions: Tally the packaged wickets and check the appropriate line.)

0-10	_____
11-15	_____
16-20	_____
21-25	_____
26-30	_____
over 30	_____

For scoring there could be a 0-5 scale, i.e., 0-10 = 0; 11-15 = 1; . . ., over 30 = 5.

5. Consumption or Quantity Used

These are measures dealing with the resources expended in performance. Often these measurements can easily be done in an unobtrusive manner.

Example 5.1:

> In order to check driving habits one might keep records on the number of replacement tires a delivery person requires each year and check it against miles driven.

Example 5.2:

> In order to check for efficient use of using electrical wire for a simulated routine telephone installation the test constructor could set standards for maximal effective use of wire; measure the amount of wire remaining after performance and check against measurement taken before performance, e.g.,
>
> Length of wire at start: _____.
> Length of wire at finish: _____.
> Length of wire used: _____.

Comment: This technique can be used for a variety of other endeavors. For example the skills test could measure the amount of computer time used, the amount of telephone usage, the amount of secretarial time used, etc.

Subjective Measurement

Subjective measures are used to classify complex processes or products into predetermined categories. The categories force the observer/scorer to make discrete decisions in regard to performances.

6. Rating Scales

Rating scales classify examinee performance on a continuum of predetermined categories.

Example 6.1:

When answering the telephone this secretary is _____.

1. overly friendly
2. courteous and professional
3. courteous but not very helpful
4. not very courteous but very helpful
5. neither courteous nor helpful

Example 6.2:

Please rate examinee performance in the four areas below by placing a "√" in the columns corresponding to your ratings.

Area	Excellent	Good	O.K.	Poor	Unacceptable	Does Not Apply
Typing letters						
Taking dictation						
Editing manuscript						
Keeping accounts accurately						

7. Forced Choice

Forced choice scoring is similar to rating scales except that the scoring is done on an "all or none" basis.

Example 7.1:

Examinee took the patient's blood pressure. (Circle one) Yes / No / N/A

or

N/A / Did Do / Did Not Do

Example 7.2:

The sales order was filled correctly. (Circle one) Yes No

Example 7.3:

For checking a series of steps a form like the one below might be used.

	Yes	No
Step 1	___	___
2	___	___
3	___	___
4	___	___

8. Checklists

Checklists are used to record the occurrence of a set of prespecified behaviors. Sometimes checklists are called "cafeteria" questions because the observer checks off what occurs from a variety of choices — none of which necessarily exclude other items.

Example 7.1:

Check all that apply to this waitress simulation

- Served water ___
- Asked if cocktails were desired ___
- Obtained cocktails from bartender ___
- Garnished cocktails ___
- Correctly returned cocktails to persons ordering them ___
- Passed out menus ___

Example 7.2:

Check all that apply to this teacher's day:

- Took attendance ___
- Collected lunch money ___
- Conducted two reading groups ___
- Had one hour of math instruction ___
- Had students at lunch on time ___

9. Attitude Scales

These measures deal with examinee attitudes toward important elements of their environment. There is a wealth of literature on constructing and using attitude scales.

Example 9.1:

I think production deadlines are _____.

a. of overriding importance.
b. very important as guidelines for production.
c. useful but not too important.
d. not particularly useful.

Example 9.2:

Reading technical literature in my field is _____.

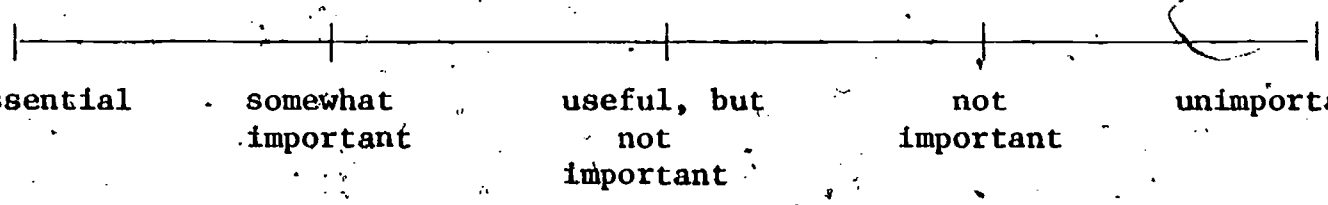
a. very important to me.
b. of some importance to me.
c. not important.

Example 9.3:

	SA	A	N	D	SD
Math classes are my favorite time during the school day.					

Example 9.4:

For the type of work I plan to do, I feel library skills are



10. Behavior Categorization

These measures deal with categorizing behaviors or the results of acts that have occurred.

Example 10.1:

Answers the telephone in a cordial manner (Check one)

very cordial	friendly	too abrupt
_____	_____	_____

Example 10.2:

Completed the sale. (Circle one) Yes Unsure No

Example 10.3:

Ability to work with subordinates. (Check one)

very effectively	effectively	somewhat effectively	ineffectively
_____	_____	_____	_____

Appendix D

Review of Non-Objective Items and Scoring

Figure D.1 Evaluation of Non-Objective Items

	Yes	No	Unsure
Criterion of Appropriateness:			
1. Is performance of this skill necessary to job success? (In other words, will there be trouble if this element is ignored?)	✓		
2. Is the element necessary for barely acceptable workers?	✓		
3. Will this element differentiate superior workers from those who are not?	✓		
4. Is it practical to expect the examinee to perform this skill at this point?	✓		
5. Has performance of this skill been deemed important vis-a-vis a validated job analysis?	✓		
Item (Task) Content:			
1. Does the task have a clear and logical beginning?	✓		
2. Does the task have a clear and logical end?	✓		
3. Does the task isolate the skills which are of interest?	✓		
4. Is the reading level appropriate for potential examinees?	✓		
5. Has the item been made excessively difficult by requiring unnecessarily exact or difficult operations?		✓	
6. Does the item give any contingencies that would unnecessarily inhibit completion?		✓	
7. Does the item present material on which the student has received instruction?	✓		
8. Is the item drawn from a validated test blueprint?	✓		
9. Can the skill be adequately performed in a given length of time?	✓		
10. If a product is to be evaluated are the expectations (specifications) delineated?	✓		

Item (Task) Structure:	Yes	No	Unsure
1. Is the task delineated in an unambiguous fashion?	✓		
2. Is the item constructed in terminology commonly used in the trade or profession?	✓		
3. Do the directions give too many cues for proper task procedures?		✓	
4. Are the task directions stated as concisely as possible?	✓		
5. Are the task directions clear?	✓		
6. Does the item clearly specify what the examinee has to do?	✓		

Response Content:			
1. Is there one clearly best way to execute the task?	✓		
2. Are there a variety of acceptable ways to execute the task?		✓	
3. Will examinees who have received training be able to select the appropriate procedure?	✓		
4. Could an examinee who has not received training execute the task?		✓	
5. Is the desired precision of performance clearly indicated in the item?	✓		

	Yes	No	Unsure
Response Structure :			
1. Are the appropriate tools or work aids available to the examinee?	✓		
2. Are the tools and work aids in good condition?	✓		
3. Is the test environment conducive to good performance?	✓		
Directions :			
1. Is the examinee informed of the fidelity which is expected?	✓		
2. Do the directions inform the examinee how responses will be scored?	✓		
3. Do the directions inform the examinee about the purposes of the test?	✓		
4. Do the directions specify whether there is only one best procedure?	✓		
5. Do the directions specify whether there are a variety of acceptable procedures?	✓		
6. Do the directions specify an appropriate amount of time which should be spent on the tasks?	✓		
7. Do the directions specify any differential weighting procedures which will be used in scoring the test?	✓		
8. Do the directions attempt to reduce examinee tension?	✓		
Post-Item (Task) Selection Considerations :			
1. Do the items represent an adequate sample of the test blueprint?	✓		
2. Are the performances appropriate to the actual job?	✓		

	Yes	No	Unsure
3. Will the sampling of different (military) procedures be confusing to the examinee?		✓	
4. Are there mechanisms to allow the examinee to proceed after poor performance on one task?	✓		

Figure D.2 Scoring a Non-Objective Test

Scoring Performance Items	Yes	No	Unsure
1. For each task has the correct procedure or the acceptable alternative been delineated?	✓		
2. Are there provisions made for partial credit where appropriate?	✓		
3. Has the manner in which performances will be ranked, rated or categorized been identified?	✓		
4. When observer judgments are used are there sample responses to represent the several possible categories?	✓		
5. Does the scoring system provide for unexpected performance?	✓		
6. Has a scoring key been prepared?	✓		
7. Have arrangements been made to have observers at the test site?	✓		
8. Are the observers likely to be personally biased due to prior interaction with the examinees?		✓	
9. Will people who have mastery in the performance area be scoring the tests?	✓	✓	
10. Will people who have mastery in the performance area be judging performances?	✓		
11. Is there adequate provision for training observers?	✓		
12. Has there been clear attempts to minimize observers making judgmental decisions?	✓		
13. Will the presence of the observer(s) effect performance?		✓	

Table B,1 Types of Tests

Characteristic	Recognition Tests	Simulation Tests	Work-Sample Tests	Project/Product Tests
Useful Situations for Application	1. large groups of examinees 2. economy is important	1. where the situation is too large and amorphous to have "real" situation 2. factors under consideration must be limited 3. where health or safety is a factor	1. when on-the-job observation is possible 2. where the work in question can be accurately observed 3. primarily used with skilled or semi-skilled workers	1. where process is not important 2. where a variety of processes are acceptable 3. when test development and administration costs are limited
Examples	1. identify parts of a diagram 2. point to specified components 3. identify functions of various components	1. role playing 2. games (computer & otherwise) 3. in-basket 4. secretarial tests	1. troubleshoot and repair 2. production output, e.g., machinist, secretary	1. artistic projects 2. sports contests 3. science fairs
Validity for Determining Proficiency	1. low for skills 2. high for knowledge	1. moderate/high for skills 2. moderate/low for knowledge	1. high for skills 2. moderate for knowledge	1. moderate/high for skills 2. moderate for knowledge
Response Modes	1. paper and pencil -multiple choice -fill in blank 2. oral 3. computer interaction	1. paper and pencil 2. computer interaction 3. oral 4. manipulative	varies; depends on actual job requirements	varies; the only response is the product

Table B.1 Types of Tests

Characteristic	Recognition Tests	Simulation Tests	Work-Sample Tests	Project/Product Tests
Scoring Modes	1. objective	1. objective, e.g., did/did not do 2. subjective, e.g., observer ratings	1. objective, e.g. output, waste, accuracy, etc. 2. subjective, e.g., rating scales, ranking, etc.	1. objective, e.g., measure tolerances, product works/does not work, amount completed, etc. 2. subjective, e.g. artistic merit
Process/Product Evaluation	does not apply	process and/or product	process and product	product
Costs	relatively inexpensive to develop, administer and score	expensive to develop, administer and score	expensive to develop costs vary to administer (often it is on-the-job time) relatively expensive to score	inexpensive to develop costs vary to administer costs vary to score
Fidelity	low	high	high	moderate/high
Useful as Instructional Device	yes	yes	no	yes
Comments		The test constructor must strive for maximum fidelity within allotted resources		