Chapter 4

PROCESS OF TEST CONSTRUCTION

In Chapter 3 we noted that the goal of most measurement in education and the social sciences is the location of individuals on a quantitative continuum with respect to a particular psychological construct. This was called "subject-centered measurement." The purpose of this chapter is to describe a process that can be followed in test construction for subject-centered measurement. The focus of this chapter is on a systematic approach to test construction that has wide applicability to various types of tests. The steps in this process can be listed as follows:

1. Identify the primary purpose(s) for which the test scores will be used
2. Identify behaviors that represent the construct or define the domain
3. Prepare a set of test specifications, delineating the proportion of items that should focus on each type of behavior identified in step 2
4. Construct an initial pool of items
5. Have items reviewed (and revise as necessary)
6. Hold preliminary item tryouts (and revise as necessary)
7. Field-test the items on a large sample representative of the examinee population for whom the test is intended
8. Determine statistical properties of item scores and, when appropriate, eliminate items that do not meet preestablished criteria
9. Design and conduct reliability and validity studies for the final form of the test
10. Develop guidelines for administration, scoring, and interpretation of the test scores (e.g., prepare norm tables, suggest recommended cutting scores or standards for performance, etc.)

The preceding list of steps probably represents the minimum effort required to ensure that test scores will have the essential technical qualities to serve as useful

measurements. Many professional test developers include additional steps in the sequence or repeat some steps several times. The sequence presented here is intended primarily as a guide to novice test developers who may undertake construction of a test for a local research or evaluation study. The main focus of this chapter is on the first six steps in the process, which are intrinsic to the initial production of a test. Procedures for executing the remaining steps (7 through 10) are the major topics of subsequent units of this text.

IDENTIFYING PURPOSES OF TEST SCORE USE

A systematic process for test development should be grounded in consideration of the basic purposes for which the test scores will be used. For example, suppose a reading expert has been commissioned to develop a test of reading comprehension for entry-level college students. Ultimately information from scores on such a test might be used for placing admissions, placement, or diagnostic decisions. Yet it is doubtful if a single test could be developed to meet all these needs optimally. As we will see in Chapter 5, a test to discriminate among examinees over a broad range of ability (or temperament) should be composed of items of medium difficulty (so that variance of examinees' scores will be maximized). On the other hand, a diagnostic test, used to identify areas of specific weaknesses for low-ability students, must contain a substantial number of items which are relatively easy for the population of examinees as a whole. Similarly the content of a test designed to assess minimum competency would probably differ from that of a test designed to select applicants for a competitive educational program. Clarifying the major purposes for which test scores will be used and establishing priorities among these probable uses greatly increases the likelihood that the final form of the test will be useful for the most important purpose it is to serve.

IDENTIFYING BEHAVIORS TO REPRESENT THE CONSTRUCT

As a general rule, the process by which psychological constructs have been translated into a specific set of test items has remained private, informal, and largely undocumented. Cronbach (1970), Roid and Haladyna (1980), and Shoemaker (1975) have discussed this subject with particular emphasis on the development of achievement tests. Typically the test developer will conceptualize or more types of behavior which are believed to manifest the construct and then simply try to "think up" items that require these behaviors to be demonstrated. Unfortunately this approach can result in omission of important areas of behavior or inclusion of areas that are relevant to the construct only in the mind of this particular test developer. Such an approach results in a highly subjective and idiosyncratic definition of
the construct. To broaden, refine, or verify the view of the construct to be measured, the test developer should engage in one or more of the following activities:

1. **Content analysis.** With this method, open-ended questions are posed to subjects about the construct of interest, and their responses are sorted into topical categories. Those topics that occur predominantly are taken as major components of the construct. For example, Jersild (1952) published results of a content analysis of compositions by children describing themselves, and the resulting categories served as the basis for generating items for two widely used inventories designed to measure children's self-concepts (Gordon, 1967; Piers and Harris, 1964).

2. **Review of research.** Those behaviors that have been most frequently studied by others are used to define the construct of interest. The test developer may use an eclectic approach or select the work of one particular theorist in specifying behavioral categories to be represented by test items.

3. **Critical incidents.** A list of behaviors is identified that characterizes extremes of the performance continuum for the construct of interest. This method is usually attributed to Flanagan (1954), who asked job supervisors to describe situations in which an employee had functioned with outstanding effectiveness or ineffectiveness and thereby generated a list of "critical behaviors" to use for rating job performance.

4. **Direct observations.** The test developer identifies the behaviors by direct observation. For example, a vocational counselor, developing an inventory to assess job-related stress in high-risk occupations, might find that actual observations of such workers on the job would help identify situations that are potential sources of emotional stress.

5. **Expert judgment.** The test developer obtains input from one or more individuals who have first-hand experience with the construct. Written questionnaires or personal interviews are used to collect information. As an illustration, a personnel psychologist who wants to develop a checklist for rating performance of staff nurses in a large hospital can survey a group of nursing supervisors to identify the types of performance that should be included.

6. **Instruction objectives.** Experts in a subject are asked to review instructional materials and develop a set of instructional objectives when an achievement test is being developed. An instructional objective specifies an observable behavior that students should be able to exhibit after completion of a course of instruction. Such objectives communicate to the item writer both the specific content on which items should focus and the nature of the tasks the examinees should be able to perform. A thorough description of the process of developing objectives is provided in most introductory texts in measurement (e.g., Brown, 1983; Mehrens and Lehmann, 1984; Popham, 1981).

**DOMAIN SAMPLING**

Many psychological constructs, such as intelligence, creativity, or moral development, are of interest primarily because of the degree to which individuals differ in the amount of that attribute. Development of a test to be used for differentiation typically involves conceptualization of major behavior components of the construct (using procedures such as those just described), production of items in these areas, and ultimately, selection of items on which the expected degree of variation in performance occurs. Items on which there is little variation are eliminated because they fail to represent the construct appropriately. Because the meaningfulness of scores on tests constructed in this fashion is derived by comparing the individual examinee's performance with the performance of others, measurements from such tests are often called *norm-referenced.*

However, in some situations, there is a need to measure examinees' performance in terms of more absolute levels of proficiency. For example, such information is often needed for achievement tests when the user wants to certify whether examinees have attained a level of minimal competency in an academic subject or to evaluate the effectiveness of an instructional program. In such situations, Ebel (1962) suggested a need for a "content standard score," which would have interpretive meaning for an examinee regardless of how others had performed on the test; and Glaser (1963) suggested that the term *criterion referenced measurement* should be applied to test scores that derive their meaning from the examinees' absolute levels of performance on a series of test items in which each set of items corresponds to a known level of proficiency on a criterion of importance. In developing tests for such purposes, the preceding methods of operationalizing the construct of interest are usually regarded as insufficient. Typically, in the development of a criterion-referenced achievement test, the test developer begins with a set of instructional objectives and then proceeds to define a domain of performance to which inferences from test scores will be made. This performance domain is called the *item domain.* An item domain is a well-defined population of items to which one or more test forms may be constructed by selection of a sample of items from this population. (This is sometimes called the *domain-sampling* approach to test construction.)

Obviously creating all possible items so that only a few could be used on a test would be economically and practically unfeasible in many cases. An alternative is to produce a set of item-domain specifications which are so structured that items written according to these specifications would be "interchangeable." Extensive reviews of methods proposed for defining item domains have been written by Millman (1974). To attain a level of minimal competency, it is not necessary to produce an item for every aspect of the domain. Thus, a sample of items is constructed that will adequately measure the construct (e.g., a sample of 10 items).
defines an entire class of items, depending on which element of the replacement set is substituted for the variable component of the item stem. An example of such an item might be

When a citrus tree displays ___(A)___, this is probably symptomatic of

1. nutritional deficiency
2. herbicide injury
3. cold weather injury
4. viral infection
5. bacterial infection

Replacement set (A) for this item would consist of a list of common leaf and bark pathological symptoms for this type of plant (e.g., yellowing leaves, splitting bark, rusty spots on leaves, black spot on leaves, scaling bark, mosaic pattern in leaves, bark cankers). The correct answer to this item would depend, varying on the element from replacement set (A) that is inserted in the item stem.

Even more-structured approaches to domain specification, which are similar to the use of item forms, involve "mapping sentences," have preidentified parts of speech, phrases, or clauses which may be systematically varied and located in designated positions of a sentence with a fixed syntactic structure. Gutmman (1969) proposed the use of mapping sentences as part of his facet theory for identifying characteristics of items for cognitive tests or attitudinal inventories that may affect response patterns. Berk (1978) and Millman (1974) discussed the application of mapping sentences to domain definition. In the assessment of reading comprehension, Bornth (1970) suggested a transformational grammar approach for generating test items based on a written passage, which is also similar to the item-forms method.

In considering the appropriate level of specificity required for domain definition in a given situation, several points should be considered. First, the meaningfulness of scores in a criterion-referenced test depends on (1) identifying a domain of behavior of practical importance and (2) constructing items in a way that allows test users to infer that performance on these items represents the performance that would be displayed on the entire item domain. Obviously there is a trade-off here. The more explicitly the domain is defined and the more individual subjectivity of different item writers is eliminated in selecting item form and content, the more confident test users can be that the items on a given test form represent the domain (in the sense that if a second test were constructed, it would closely resemble the first). Yet as the domain is restricted, the extent to which it is likely to be considered "practically important" may be reduced. Being able to generalize with confidence to a domain that encompasses a very limited sphere of behavior may be appropriate for some measurements but questionable in others.

Second, a related issue is whether depending upon item specifications to reduce the degree of subjectivity involved in selection of item form and content simply displaces subjective decision making from the point of item-writing to the point at
which item specifications are written. The consequences of this shift require careful consideration.

Finally, if a particular item specification results in consistent production of items that are technically flawed or susceptible to response set, a test could consist of a much greater proportion of faulty items than might occur if item writers had not been constrained to follow specifications. It is important, therefore, to have the item specifications carefully reviewed prior to their use and to develop several items according to each item specification on a pilot basis so that potential flaws in the specification can be detected.

PREPARING TEST SPECIFICATIONS

Once a set of objectives, item specifications, or other categories of behavior have been chosen, the test developer needs to formulate a plan for deciding the relative emphasis that each of these components should receive on the test. Specifically there should be a balance of items so that different components of the construct are represented in proportion to the test developer's perception of their importance. If an item domain has been defined through a set of item specifications, the issue becomes, How many items on the test should represent proportionally each item specification? Without such a plan, subsequent forms of a test produced from the same set of item specifications might differ substantially if the writers emphasized some item specifications more heavily than others and the areas of emphasis changed from form to form. In the more common case, where no item specifications have been prepared, the development of test specifications usually requires the test developer to attend to two orthogonal properties of items—the substantive content and the cognitive process or operation which the examinee apparently must employ to carry out the item task. This requirement is especially critical in the development of achievement tests, as shown in the example that follows.

Consider the following objectives from the same instructional unit in a plane geometry course:

A. Define basic terms related to circles (e.g., radius, diameter, central angle).
B. Compute areas, distances, circumferences, and angle measures by using properties of circles.

Obviously the first objective requires primarily recall of memorized material, whereas the second requires both knowledge of these concepts and application of principles defining relationships of two or more concepts. Thus to simply specify the content to be covered by the items of a test without also indicating the levels or types of cognitive operations that should be represented does not provide adequate guidance for test development.

Several authors have suggested hierarchical systems for categorizing cognitive operations; these are useful in developing test specifications. Probably the best known of these is the taxonomy by Bloom (1956) consisting of the following major categories:

1. Knowledge—recall of factual material in similar form to that in which it was presented during instruction; an objective at this level might entail naming the capital cities of given states.
2. Comprehension—translation, interpretation, or extrapolation of a concept into somewhat different form than originally practiced or presented; an objective at this level might require recognizing nouns in sentences that have not been used in class examples or giving an example of a prime number other than those given in the text or class examples.
3. Application—solving new problems through the use of familiar principles or generalizations, for example, calculating the resistance of an electrical conductor by using Ohm’s law when no reference to the law is made in the problem statement.
4. Analysis—breaking down a communication or problem into its component elements by using a process that requires recognition of multiple elements, relationships among these elements, and/or organizational principles, for example, making an identification of the genus of a new specimen plant based upon its leaf and flower characteristics.
5. Synthesis—combining elements into a whole by using an original structure or solving a problem that requires combination of several principles sequentially in a novel situation, for example, writing a computer program to perform a calculation on a set of data records using input, output, loop, and logical transfer statements in an efficient sequence of execution.
6. Evaluation—employment of internal (self-generated) or external criteria for making critical judgments in terms of accuracy, consistency of logic, or artistic or philosophical point of view, for example, writing a critical review of a journal article describing empirical research in social or personality psychology.

Bloom, Hastings, and Madaus (1971) provided additional explanation and examples of these categories. Additional systems for categorizing processes required on optimal performance tests were described by Ebel (1972), who suggested the processes of knowledge and use of terminology; recall of factual material; use of generalizations, explanation, calculation, and prediction of outcomes; and recommendations for action. Thorndike and Hagen (1977) detailed the categories of recognizing terms and vocabulary; identifying facts; identifying principles, concepts, and generalizations; evaluating information; and applying principles to novel situations. It is not uncommon for test developers to select only a few processes from any of these taxonomies or perhaps to combine categories from one or more taxonomies to suit their specific needs.

Because of the great variation in the level of cognitive processes that may occur in items covering even a single topic, test specifications should provide some guidance to item writers in terms of the levels of operation that should be represented. Test specifications should also indicate the balance that should be maintained in trying to tap these different processes for the test as a whole. One useful structure for test specifications that takes into account both content and process is known as
a table of specifications. The table is basically a two-way grid with major content areas listed on one margin and cognitive processes on the other. An illustrative table of specifications for a subtest of a teacher certification examination is shown in Figure 4.2. (This example illustrates that process levels may be selected from more than one taxonomy.) The number in each cell is an arbitrary value weight indicating the relative emphasis in the examination that the test developer wishes to place on the content and process associated with that cell. In an achievement test, these cell weights usually reflect the amount of time devoted to mastery of material at that level during the course of instruction. The total of these cell weights should equal 100. Each weight indicates the percentage of items or points on the test that should be devoted to coverage of this content and process. Such weights determine how many items must be written to represent the cells adequately in relation to the domain of interest. In Figure 4.2, 1% of the test items should measure recall of classroom management principles, and 10% should measure application of classroom management principles. The table of specifications also provides a convenient way to describe the test to potential users. By summing the percentages in each column, we can see that 7% of the test items measure knowledge-level skills, 62% measure application-level skills, and 31% measure problem-solving skills. Summing the percentages in each row yields the total percentage of items covering each category respectively. When the test developer has decided on the total number of items, the number of items that must be written for any cell to meet the specifications equals the cell weight times \( K \), where \( K \) equals the total number of items on the test.

In the construction of attitude inventories or observational performance measures, the test developer may wish to use a table of specifications that involves behaviors from the affective domain. Krathwohl, Bloom, and Masia (1964) have presented a taxonomy which may be useful for this purpose. Detailed description of the five categories they proposed and examples of their use in test development may be found in Bloom, Hastings, and Madaus (1971).

ITEM CONSTRUCTION

Lindquist (1936) characterized the test developer's task as requiring two major types of decisions—what to measure and how to measure it. During item construction the latter type of decision must be addressed. Developing a pool of items to measure a construct entails the following activities:

1. Selecting an appropriate item format
2. Verifying that the proposed format is feasible for the intended examinees
3. Selecting and training the item writers
4. Writing the items
5. Monitoring the progress of the item writers and the quality of the items

If item writers will work from item forms or item specifications, the structure and format of the items may already be determined. If item specifications are not being used, it is still important for decisions about item format to be made at the outset of the item-construction phase rather than to be left to the idiosyncratic tastes of individual writers. In deciding on a common format, the test developer may wish to review similar instruments in the field and study reports of their development. The opinions of experts may also be helpful in deciding such matters as whether the examinees are sufficiently literate to take group-administered pencil-and-paper tests, whether company employees would take the time to write out responses to open-ended questions, or whether third-graders can distinguish among five points on an agree-disagree continuum. At times it may be necessary to collect data in a small-scale study, using a few prototype items to answer such questions, before producing a large number of items in a particular format.

Once an appropriate item format has been chosen, the test developer should
review standard sources on item writing to glean suggestions on writing items of this type. A list of guidelines should be prepared and distributed to the item writers, particularly if nonprofessional writers are to be employed.

**Item Formats for Optimal Performance Tests**

For optimal performance tests (typified by aptitude or achievement tests), a wide variety of item formats may be considered. Popham (1981) has divided these formats into two major categories—those that require the examinee to generate the response (e.g., essay or short-answer questions) and those that provide two or more possible responses and require the examinee to make a selection. Because the latter can be scored with little subjectivity, they are often called objective test items. The three most widely used objective formats are:

1. **Alternate choice**—a statement (or question) and two possible responses (true-false or yes-no).
   
   Example: The most probable cause of "sooty," black coating on the leaves of a citrus tree, is insect infestation. True False

2. **Multiple choice**—a stem in which the question or problem is posed; a correct response or keyed response; and two or more incorrect responses, which are called "foils.
   
   Example: A citrus tree has a number of leaves coated with a black, "sooty" substance. This is probably caused by
   a. Herbicide damage
   b. Bacterial infection
   c. Insect infestation
   d. Nutritional deficiency

3. **Matching**—a statement of the principle to be used in relating the objects from two separate lists, a list of premises or stimuli, and a list of responses.
   
   Example: For each symptom listed in the left column, record the letter from the right column which identifies the most probable cause.

<table>
<thead>
<tr>
<th>Symptom</th>
<th>Probable Cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Yellow leaves</td>
<td>A. Cold weather injury</td>
</tr>
<tr>
<td>2. Splitting bark</td>
<td>B. Nitrogen deficiency</td>
</tr>
<tr>
<td>3. Rusty spots on leaves</td>
<td>C. Fungal infection</td>
</tr>
<tr>
<td>4. Mottled leaves</td>
<td>D. Herbicide damage</td>
</tr>
<tr>
<td></td>
<td>E. Viral infection</td>
</tr>
</tbody>
</table>

Most introductory texts in testing and measurement devote several chapters to suggestions for writing each of these formats. For example, the reader might wish to consult Brown (1983), Mehrens and Lehmann (1984), Popham (1981), or Sax (1980). An important point that is common to all objective formats, however, is that all responses should appear logically reasonable to an examinee who does not have the knowledge or skill that the item was designed to test. For example, in a multiple-choice item, a well-written item stem accompanied by the correct response and three absurd foils is unlikely to measure examinees' knowledge in the domain intended; most examinees would be able to eliminate all but the correct choice by using only common sense, and as a result, the usefulness of the item would be reduced. To avoid this problem, foils for multiple-choice items are often constructed from common misconceptions, misinterpretations, or answers that could result from errors of computation. The point here is that in selecting item writers, the test developer should remember that construction of a good item requires not only knowledge of the material but also familiarity with the examinee population to ensure that incorrect choices will appear plausible.

As a general rule, test developers are advised to give careful thought to selecting an item format appropriate to the needs of the examinees and to avoid novel or untried formats without having a sound rationale for their use. Recently, however, some authors have called for moving beyond the use of the highly popular multiple-choice format, although the direction to take has generated strongly contrasting points of view. On the one hand, Ebel (1982) advocated broader use of more highly structured alternate-choice items—for example, "The density of ice is (1) greater (2) less than that of water." His point was that in many cases the domain of knowledge that is to be sampled could actually be expressed as a series of precisely stated functional relationships or principles and that each of these can form the basis of a highly structured item. Although it would appear that each item can test only a limited "piece" of information, Ebel argued that (1) a large number of such items can be asked in a relatively short amount of time, thus permitting more thorough sampling of the content domain; (2) performance on such items is less subject to the influence of extraneous factors, such as the higher level of reading ability that may be required by more complex formats; and (3) such items have greater conceptual clarity for examinees about the nature of what is being asked (or the underlying principle being tested). On the other hand, Frederiksen (1981, p. 19) suggested that in the pursuit of greater scoring efficiency, test developers may rely too heavily on highly structured formats and that this interest in what can be easily measured may overshadow considerations about what should be measured. This may leave "many important abilities untested and untaught." As one alternative, Frederiksen suggested the development of unstructured multiple-choice problem-solving tasks. Recent developmental work on such a test has been described by Ward, Carson, and Woitschlagter (1983). At this time, the issue of whether items covering the same material but using such contrasting formats actually measure the same trait has not been conclusively resolved, but the issue seems likely to spawn additional theoretical and empirical investigation.

**Item Formats for Inventories**

Three popular item formats for attitude and personality inventories are the dichotomous Agree-Disagree format, the Likert format, and the bipolar adjective checklist. Each of these will be briefly described.

The dichotomous Agree-Disagree format typically consists of a simple declara-
scale. The median of the judges' ratings is the item weight. Furthermore, the items to appear on the scale are chosen from the initial large set of items on the basis of a statistic such as $Q$, the semi-interquartile range:

$$Q = \frac{X_{75} - X_{25}}{2}$$

where $X_{75}$ is the numeric value corresponding to the 75th percentile rank, and $X_{25}$ is the numeric value corresponding to the 25th percentile rank in the distribution of ratings. Smaller values of $Q$ indicate that judges were in fairly close agreement about the strength of sentiment expressed by the item. Items with smaller $Q$ values are, therefore, favored in item selection although the test developer still tries to include some items from each category. When the scale is later administered to a group of respondents, each time a respondent endorses a statement, the weight value for this item is added to the respondent's total score. This total score is then divided by the number of items endorsed to obtain the average weight of the statements endorsed. The average weight is used for comparing or describing respondents' attitudes.

The second widely used item format for inventories was suggested by Likert (1932). This method requires writing a collection of statements, each of which is clearly positive or clearly negative with respect to the construct of interest. Statements neutral in affect are not included. The respondent reads each statement and selects a response from a five-point continuum ranging from "Strongly Agree" to "Strongly Disagree," as shown in the following example:

Children should obey their parents without question.

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

The statements are not weighted or scaled prior to their administration to respondents, but respondents use the response continuum to indicate the degree of strength of their endorsement. To score items with a graded response continuum, 1 point is assigned to the response showing the lowest level of positive sentiment toward the construct; 2 points are assigned to the response showing the next highest level; 3 points to the response indicating the next highest level; and so forth. If we again assume that the test developer wants respondents with more permissive attitudes to receive higher scores, the three items from the preceding example would have scoring weights as follows:

<table>
<thead>
<tr>
<th>Statement</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Children should obey their parents without question.</td>
<td>1.00 2.00 3.00 4.00 5.00</td>
</tr>
<tr>
<td>2. Children today need stronger discipline at home.</td>
<td>1.00 2.00 3.00 4.00 5.00</td>
</tr>
<tr>
<td>3. Most children with behavior problems are rebelling against too much parental control.</td>
<td>1.00 2.00 3.00 4.00 5.00</td>
</tr>
</tbody>
</table>

The data resulting from this judgment task are used for two purposes—selection of items for the final version of the scale and assignment of weights for scoring the
A respondent's total score is the sum of the points associated with the response given to each of the items. The following are some general guidelines that may be helpful in writing and reviewing inventory items for either Likert or agree-disagree formats:

1. Put statements or questions in the present tense.
2. Do not use statements that are factual or capable of being interpreted as factual.
3. Avoid statements that can have more than one interpretation.
4. Avoid statements that are likely to be endorsed by almost everyone or almost no one.
5. Try to have an almost equal number of statements expressing positive and negative feelings.
6. Statements should be short, rarely exceeding 20 words.
7. Each statement should be a proper grammatical sentence.
8. Statements containing universals such as all, always, none, and never often introduce ambiguity and should be avoided.
9. Avoid use of indefinite qualifiers such as only, just, merely, many, few, or seldom.
10. Whenever possible, statements should be in simple sentences rather than complex or compound sentences. Avoid statements that contain "if" or "because" clauses.
11. Use vocabulary that can be understood easily by the respondents.
12. Avoid use of negatives (e.g., not, none, never).

Recent references on the production of attitude statements include those by Anderson (1981); Dawes (1972); and Udinsky, Osterling, and Lynch (1981).

Another popular format sometimes used in inventory construction is the bipolar adjective pair. The origin of this format is usually attributed to Osgood, Suci, and Tannenbaum (1957), who proposed its use in the study of the semantic meaning of psychological constructs. The construct of interest is listed at the top, followed by a pair of adjectives which should represent opposite poles of a single continuum. Typically a five- or seven-point continuum is presented between the adjectives, and respondents are instructed to mark the spot on the continuum which most closely reflects their feeling. For example, in attempting to assess teachers' attitudes toward working with mentally retarded children, the following format might be used:

<table>
<thead>
<tr>
<th>Mentally Retarded Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretty</td>
</tr>
<tr>
<td>Happy</td>
</tr>
<tr>
<td>Dirty</td>
</tr>
<tr>
<td>Ugly</td>
</tr>
<tr>
<td>Sad</td>
</tr>
<tr>
<td>Clean</td>
</tr>
</tbody>
</table>

Osgood and his colleagues, after investigating a large number of diverse adjective pairs, determined that most of them could be grouped into one of three dimensions (evaluation, potency, or activity) underlying the semantic meaning of verbal constructs. They suggested a method for analyzing and interpreting responses to these items and called their instrument a semantic differential. The test developer who generates a unique set of bipolar adjectives for a specific assessment, without using these pretested pairs, must assume responsibility for determining how the responses should be scored and interpreted. In this situation, the item format should probably be described as a bipolar adjective checklist rather than a semantic differential measure.

In the development of inventory items, it is important to be aware of response sets that may affect individual examinees' behavior. Response set may be defined as the tendency of an examinee to respond in a certain way to a particular item format regardless of content. Guilford (1954) provided a classic identification of several common types of response sets and offered suggestions for reducing or controlling their effects. Two response sets that often affect inventory scores are acquiescence (the tendency to agree with a statement regardless of its content) and differential individual interpretations of indefinite qualifiers (e.g., some and often). Adherence to the previous writing suggestions can reduce the likelihood that items will be susceptible to response sets. Other issues related to construction of inventory items include the effects of using different types of anchor points on the response continuum (see Frisch and Brandenburg, 1979; Lam and Klockars, 1982) or different numbers of response alternatives (see Masters, 1974; Velicer and Stevenson, 1978). Such studies again illustrate that decisions about how to measure behavior may have an impact ultimately on what has been measured.

**ITEM REVIEW**

As test items are drafted, it is advisable for the test developer to ask qualified colleagues to review them informally for accuracy, wording, grammar, ambiguity, and other technical flaws. "Problem" items can then be revised as necessary. In addition, once the items have been written, they should be subjected to more formal review on an item-by-item basis. Important aspects of item construction which should be considered include:

1. Accuracy
2. Appropriateness or relevance to test specifications
3. Technical item-construction flaws
4. Grammar
5. Offensiveness or appearance of "bias"
6. Level of readability

Different types of expertise are required on the item review panel. For example, experts in the subject matter are best qualified to certify that the items are clearly stated and correctly keyed. They are also qualified to judge whether items are appropriate for the test specifications or item specifications. (More detailed explanation of this process is given in Chapter 10 on content validation.) Some general expertise in measurement and test construction is important for the reviewer(s) who must certify that items are free from construction flaws. For example, if items are in
multiple-choice format, the technical expert should look for common flaws affiliated with this particular format (e.g., making the correct alternative longer than the others). The texts mentioned in the section on item construction can be used to identify common flaws or problems most likely to occur with a particular format. In addition, the measurement expert should identify the types of response sets most likely to affect scores on the instrument being developed and consider the susceptibility of each item to these types of response sets. Naturally, every item on a test should be free of grammatical errors, including spelling errors. Particularly, flaws in punctuation or unwieldy sentence construction may result in misinterpretation. All item reviewers should be instructed to look for grammatical errors.

Also, one or more members of the review panel should have expert familiarity with the population for whom the test is intended. These reviewers should consider whether content might be construed as offensive or seemingly "biased" toward any particular subgroup, perhaps by use of undesirable cultural stereotypes or content that is unfamiliar to certain subgroups (when this content is unrelated to the construct or knowledge domain being tested). For example, a math word problem requiring calculation of simple interest on the sales price of an object should involve an object that is appropriate for purchase by most examinees (or their families) rather than an exclusive luxury item that would more likely be purchased by upper-middle-class or wealthy families only.

For items to be used with children or adolescents, the readability level of the items should be considered when the test is being developed to measure something other than reading skills. Many standard methods for assessing readability of written passages are not appropriate for test items because of their length and structure, but one promising procedure has been described by Ironson and her associates (1984).

Item review can be carried out either before or after preliminary item tryouts. The choice of sequence is made on the basis of convenience and economy. If expert review is conducted before tryouts, the time is not costly, item review conducted before tryouts so that time in tryouts will not be wasted on faulty or biased items. On the other hand, after item tryouts many items will inevitably be revised or reworded. This creates the necessity for additional item review by the expert panel. Thus if substantial costs or effort are involved in assembling the review panel, many test developers choose to defer this activity until after preliminary item tryouts and subsequent revisions. If results of the item review are to be reported as evidence of content validity, it is especially important for the review panel to examine items in their final form.

PRELIMINARY ITEM TRYOUTS

Before the test developer has printed items in final form for a field test, it is a good idea to try out the items on a small sample of examinees. If only a limited number of subjects are available (as when participants in an experimental program are the examinees of interest), most of these subjects must be reserved for later field trials of the items. In such cases, it might be necessary to use as few as 15 to 30 subjects for the preliminary item tryouts. Items developed for commercial use may be tested on samples as large as 100 to 200. If a large number of items have been developed and testing time is limited, it is possible to administer subsets of items to different groups of examinees. Preliminary item tryouts are fairly informal, and the test developer should use this opportunity to observe examinees' reactions during testing, noting such behaviors as long pauses, scribbling, or answer-changing, which may indicate confusion about particular items. After the testing session, a "debriefing" should take place in which examinees are invited to comment on each item and offer suggestions for possible improvements.

Examination of descriptive statistics for the response distribution to each item is also recommended. This will enable the test developer to obtain a rough idea of whether the items seem to be at the appropriate level of difficulty for the group as a whole and whether there is sufficient variation in the responses to justify proceeding into a larger-scale field test. It is important to recognize that although the final decision about which items to retain and which to eliminate are made on the basis of the large-scale field test, items are often revised extensively after reviewing the results of preliminary tryouts.

THE NEXT STEPS

After items have been through preliminary tryouts (and possibly subsequent revisions), they are ready for a full-fledged field test. The field test typically involves the administration of the items in their final draft form to a large sample of examinees representative of those for whom the test is designed. Statistical properties of the item scores are examined through a variety of procedures, known as item analysis. Designing item field-test studies and conducting appropriate analyses of item response data is the theme of Unit IV of this text. Typically, when norm-referenced interpretations are to be made, the test developer will use results of item analysis to "cull" those items that do not appear to function as intended. As we shall learn in Chapter 14, however, there is lack of agreement about using item analysis results to eliminate items from criterion-referenced tests.

Once a final form of the test is assembled, it is incumbent on the test developer to undertake studies of the test scores' reliability and validity. Theories and practices relevant to conducting studies of testscore reliability are the topics of Chapters 6 through 9 (Unit II); validation procedures are the focus of Chapters 10 through 12 (Unit III). Suggestions for development of scoring procedures, setting standards, and providing normative data to aid in test score interpretation are presented in Unit V.

SUMMARY

In this chapter ten steps in a systematic process of test development were identified. First, the purpose(s) for which test scores are most likely to be used must be clearly
specified. The most appropriate approach to test development must be chosen in light of these purposes. Second, two possible approaches to test development were described: development of a test for differentiation among individuals on a given construct, and development of a test designed to provide scores that describe more absolute levels of proficiency in a given content area. Useful aids in identifying behaviors which may typify a construct include content analysis, review of research, critical incidents, direct observations, and instructional objectives. In the development of achievement tests for which criterion-referenced measurements are desired, the goal of test developers is to establish a well-defined item domain so that test users may infer from test scores on a sample of items proficiency with respect to the entire item domain. Consequently the item domain may be established through instructional objectives, item specifications, and item forms or item algorithms.

The third step in test construction is development of a table of specifications, which delineates the proportion of items on the test that should focus on various content and process categories relevant to the construct of interest (or item domain).

Several taxonomies for categorizing behavior in the cognitive domain were presented. The most well known is the one proposed by Bloom (1956). In addition, a taxonomy proposed for behavior in the affective domain has been described by Krathwohl, Bloom, and Masia (1964).

Fourth, item construction entails selection of an appropriate item format—verifying that the format is feasible for use with the intended population—selection and training of item writers, item writing, and monitoring the progress of the item writers. Three popular objective formats for optimal performance tests are the alternate-choice, matching, and multiple-choice. For inventories, three widely used formats are the agree-disagree format often used with Thurstone scaling, the Likert format, and the bipolar adjective checklist. Each of these item formats was briefly described.

Fifth, items should also be subjected to review by a panel of experts who should consider such aspects as accuracy, relevance to test specifications, technical quality, grammar, potential for offensiveness or appearance of cultural or gender bias, and readability. In addition, items should be administered to small groups of examinees for informal preliminary tryouts to ascertain that examinees can follow the instructions associated with the format, to obtain estimates of the time required to take the test, and to identify items that are poorly written or ambiguous. Revisions indicated from the item review or preliminary item tryouts should be made before proceeding with formal field-testing of the items on a large sample. The final steps in test development include field-testing and item analysis, reliability and validity studies, and establishing guides for test score interpretation. These procedures are covered in subsequent units of this text.

Exercises
1. Review objectives A through L, based on the contents of Chapter 2, and classify them into cells of the table of specifications that follows.
Chapter 5

TEST SCORES AS COMPOSITES

A test battery is a collection of two or more separate tests designed to be administered to the same examinees. A separate score is computed for each test on the battery. A well-known example of a test battery is the Graduate Record Examination battery, from which separate verbal, quantitative, and analytic subtest scores are obtained. A composite test score is a total test score created by summing two or more subtest scores. To interpret correctly the scores of such a composite, it is important to understand how the statistical properties of subtest scores influence those of the composite score. Even more important, however, test developers must recognize that every test score is a composite. An item score can be defined as the number of points assigned to an individual's response to a given item. In scoring a test, the total test score is usually determined by summing the item scores. Thus each item could be considered a very short subtest, and item scores can be considered as "mini-test" scores. In this sense whenever a test score is created by summing the points awarded to an examinee on each item, that total test score is a composite.

In this chapter we will consider the statistical operations that can be performed on item scores and the relationship of the statistical properties of item scores to those of the total test score. An understanding of the relationship is crucial for effective test development because items are the building blocks of which tests are made. A test can have no property that is not a function of the items that comprise it. The test developer who seeks to create a test of sound quality must have information about the score distribution of each item and its relationship to other items in the test. In Chapter 2 we considered how the concepts of mean, variance, standard deviation, and correlation could be applied to describing the distribution of test scores. Now we will consider how these statistics can be applied to item scores and how the item statistics affect distributional properties of total test scores.
INTRODUCTION
TO CLASSICAL
AND MODERN
TEST THEORY

Linda Crocker
James Algina

University of Florida