CHAPTER 8

Test Development

All tests are not created equal. The creation of a good test is not a matter of chance. It is the product of the thoughtful and sound application of established principles of *test development.* In this context, **test development** is an umbrella term for all that goes into the process of creating a test.

In this chapter, we introduce the basics of test development and examine in detail the processes by which tests are assembled. We explore, for example, ways that test items are written, and ultimately selected for use. Although we focus on tests of the published, standardized variety, much of what we have to say also applies to custom-made tests such as those created by teachers, researchers, and employers.

The process of developing a test occurs in five stages:

1. test conceptualization;
2. test construction;
3. test tryout;
4. item analysis;
5. test revision.

Once the idea for a test is conceived (**test conceptualization**), *test construction* begins. As we are using this term, **test construction** is a stage in the process of test development that entails writing test items (or re-writing or revising existing items), as well as formatting items, setting scoring rules, and otherwise designing and building a test. Once a preliminary form of the test has been developed, it is administered to a representative sample of testtakers under conditions that simulate the conditions that the final version of the test will be administered under (**test tryout**). The data from the tryout will be collected and testtakers' performance on the test as a whole and on each item will be analyzed. Statistical procedures, referred to as *item analysis,* are employed to assist in making judgments about which items are good as they are, which items need to be revised, and which items should be discarded. The analysis of the test's items may include analyses of item reliability, item validity, and item discrimination. Depending on the type of test, item-difficulty level may be analyzed as well.

Next in the sequence of events in test development is *test revision.* Here, **test revision** refers to action taken to modify a test's content or format for the purpose of improving the test's effectiveness as a tool of measurement. This action is usually based on item analyses, as well as related information derived from the test tryout. The revised version of the test will then be tried out on a new sample of testtakers. After the results are

JUST THINK . . .

Can you think of a classic psychological test from the past that has never undergone test tryout, item analysis, or revision? What about so-called psychological tests found on the Internet?

analyzed the test will be further revised if necessary—and so it goes (see Figure 8–1). Although the test development process described is fairly typical today, let's note that there are many exceptions to it, both with regard to tests developed in the past, and some contemporary tests. Some tests are conceived of and constructed but neither tried-out, nor item-analyzed, nor revised.

Test Conceptualization

The beginnings of any published test can probably be traced to thoughts—self-talk, in behavioral terms. The test developer says to himself or herself something like: "There ought to be a test designed to measure [fill in the blank] in [such and such] way." The stimulus for such a thought could be almost anything. A review of the available literature on existing tests designed to measure a particular construct might indicate that such tests leave much to be desired in psychometric soundness. An emerging social phenomenon or pattern of behavior might serve as the stimulus for the development of a new test. The analogy with medicine is straightforward: Once a new disease comes to the attention of medical researchers, they attempt to develop diagnostic tests to assess its presence or absence as well as the severity of its manifestations in the body.

The development of a new test may be in response to a need to assess mastery in an emerging occupation or profession. For example, new tests may be developed to assess mastery in fields such as high-definition electronics, environmental engineering, and wireless communications.

In recent years, measurement interest related to aspects of the LGBT (lesbian, gay, bi-sexual, and transgender) experience has increased. The present authors propose that in the interest of comprehensive inclusion, an "A" should be added to the end of "LGBT" so that this term is routinely abbreviated as "LGBTA." The additional "A" would acknowledge the existence of asexuality as a sexual orientation or preference.

Asexuality may be defined as *a sexual orientation characterized by a long-term lack of interest in a sexual relationship with anyone or anything.* Given that some research is conducted with persons claiming to be asexual, and given that asexual individuals must be selected-in or selected-out to participate in such research, Yule et al. (2015) perceived a need for a reliable and valid test to measure asexuality. Read about their efforts to develop and validate their rather novel test in this chapter's *Close-Up.*

JUST THINK . . .

What is a "hot topic" today that developers of psychological tests should be working on? What aspects of this topic might be explored by means of a psychological test?

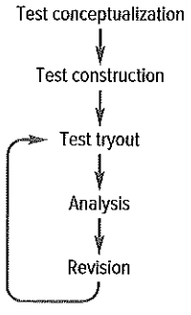

Figure 8–1
The Test Development Process

CLOSE-UP

Creating and Validating a Test of Asexuality*

In general, and with some variation according to the source, **human asexuality** may be defined as an absence of sexual attraction to anyone at all. Estimates suggest that approximately 1% of the population might be asexual (Bogaert, 2004). Although the concept of asexuality was first introduced by Alfred Kinsey in 1948, it is only in the past decade that it has received any substantial academic attention. Scholars are grappling with how best to conceptualize asexuality. For some, asexuality is thought of as itself, a sexual orientation (Berkey et al., 1990; Bogaert, 2004; Brotto & Yule, 2011; Brotto et al., 2010; Storms, 1978; Yule et al., 2014). Others view asexuality more as a mental health issue, a paraphilia, or human sexual dysfunction (see Bogaert, 2012, 2015).

More research on human asexuality would be helpful. However, researchers who design projects to explore human asexuality face the challenge of finding qualified subjects. Perhaps the best source of asexual research subjects has been an online organization called "AVEN" (an acronym for the Asexuality and Visibility Education Network). Located at *asexuality.org,* this organization had some 120,000 members at the time of this writing (in May, 2016). But while the convenience of these group members as a recruitment source is obvious, there are also limitations inherent to exclusively recruiting research participants from a single online community. For example, asexual individuals who do not belong to AVEN are systematically excluded from such research. It may well be that those unaffiliated asexual individuals differ from AVEN members in significant ways. For example, these individuals may have lived their lives devoid of any sexual attraction, but have never construed themselves to be "asexual." On the other hand, persons belonging to AVEN may be a unique group within the asexual population, as they have not only acknowledged their asexuality as an identity, but actively sought out affiliation with other like-minded individuals. Clearly, an alternative recruitment procedure is needed. Simply relying on membership in AVEN as a credential of asexuality is flawed. What is needed is a validated measure to screen for human asexuality.

In response to this need for a test designed to screen for human asexuality, the Asexuality Identification Scale (AIS) was developed (Yule et al., 2015). The AIS is a 12-item, sex- and gender-neutral, self-report measure of asexuality. The AIS was developed in a series of stages. Stage 1 included development and administration of eight open-ended questions to sexual ($n = 70$) and asexual ($n = 139$) individuals. These subjects were selected for participation in the study through online channels (e.g., AVEN, Craigslist, and Facebook). Subjects responded in writing to a series of questions focused on definitions of asexuality, sexual attraction, sexual desire, and romantic attraction. There were no space limitations, and participants were encouraged to answer in as much or as little detail as they wished. Participant responses were examined to identify prevalent themes, and this information was used to generate 111 multiple-choice items. In Stage 2, these 111 items were administered to another group of asexual ($n = 165$) and sexual ($n = 752$) participants. Subjects in this phase of the test development process were selected for participation through a variety of online websites, and also through our university's human subjects pool. The resulting data were then factor- and item-analyzed in order to determine which items should be retained. The decision to retain an item was made on the basis of our judgment as to which items best differentiated asexual from sexual participants. Thirty-seven items were selected based on the results of this item selection process. In Stage 3, these 37 items were administered to another group of asexual ($n = 316$) and sexual ($n = 926$) participants. Here, subjects were selected through the same means as in Stage 2, but also through websites that host psychological online studies. As in Stage 2, the items were analyzed for the purpose of selecting those items that best loaded on the asexual versus the sexual factors. Of the 37 original items subjected to item analysis, 12 items were retained, and 25 were discarded.

In order to determine construct validity, psychometric validation on the 12-item AIS was conducted using data from the same participants in Stage 3. Known-groups validity was established as the AIS total score showed excellent ability to distinguish between asexual and sexual subjects. Specifically, a cut-off score of 40/60 was found to identify 93% of self-identified asexual individuals, while excluding 95% of sexual individuals. In order to assess whether the measure was useful over and above already-available measures of sexual orientation, we compared the AIS to an adaptation of a previously established measure of sexual orientation (Klein Scale; Klein & Sepekoff, 1985). Incremental validity was established, as the AIS showed only moderate correlations with the Klein Scale, suggesting that the AIS is a better predictor of asexuality compared to an existing measure. To determine whether the AIS correlates with a construct that is thought to be highly related to asexuality (or, lack of sexual desire), convergent validity was assessed by correlating total AIS

(continued)

*This *Close-Up* was guest-authored by Morag A. Yule and Lori A. Brotto, both of the Department of Obstetrics & Gynaecology of the University of British Columbia.

Creating and Validating a Test of Asexuality *(continued)*

scores with scores on the Sexual Desire Inventory (SDI; Spector et al., 1996). As we expected, the AIS correlated only weakly with Solitary Desire subscale of the SDI, while the Dyadic Desire subscale of the SDI had a moderate negative correlation with the AIS. Finally, we conducted discriminant validity analyses by comparing the AIS with the Childhood Trauma Questionnaire (CTQ; Bernstein et al., 1994; Bernstein & Fink, 1998), the Short-Form Inventory of Interpersonal Problems-Circumplex scales (IIP-SC; Soldz et al., 1995), and the Big-Five Inventory (BFI; John et al., 1991; John et al., 2008; John & Srivastava, 1999) in order to determine whether the AIS was actually tapping into negative sexual experiences or personality traits. Discriminant validity was established, as the AIS was not significantly correlated with scores on the CTQ, IIP-SC, or the BFI.

Sexual and asexual participants significantly differed in their AIS total scores with a large effect size. Further, the AIS passed tests of known-groups, incremental, convergent, and discriminant validity. This suggests that the AIS is a useful tool for identifying asexuality, and could be used in future research to identify individuals with a lack of sexual attraction. We believe that respondents need not be self-identified as asexual in order to be selected as asexual on the AIS. Research suggests that the AIS will identify as asexual the individual who exhibits characteristics of a lifelong lack of sexual attraction in the absence of personal distress. It is our hope that the AIS will allow for recruitment of more representative samples of the asexuality population, and contribute toward a growing body of research on this topic.

Used with permission of Morag A. Yule and Lori A. Brotto.

Some Preliminary Questions

Regardless of the stimulus for developing the new test, a number of questions immediately confront the prospective test developer.

- *What is the test designed to measure?* This is a deceptively simple question. Its answer is closely linked to how the test developer defines the construct being measured and how that definition is the same as or different from other tests purporting to measure the same construct.
- *What is the objective of the test?* In the service of what goal will the test be employed? In what way or ways is the objective of this test the same as or different from other tests with similar goals? What real-world behaviors would be anticipated to correlate with testtaker responses?
- *Is there a need for this test?* Are there any other tests purporting to measure the same thing? In what ways will the new test be better than or different from existing ones? Will there be more compelling evidence for its reliability or validity? Will it be more comprehensive? Will it take less time to administer? In what ways would this test *not* be better than existing tests?
- *Who will use this test?* Clinicians? Educators? Others? For what purpose or purposes would this test be used?
- *Who will take this test?* Who is this test for? Who needs to take it? Who would find it desirable to take it? For what age range of testtakers is the test designed? What reading level is required of a testtaker? What cultural factors might affect testtaker response?
- *What content will the test cover?* Why should it cover this content? Is this coverage different from the content coverage of existing tests with the same or similar objectives? How and why is the content area different? To what extent is this content culture-specific?
- *How will the test be administered?* Individually or in groups? Is it amenable to both group and individual administration? What differences will exist between individual and

group administrations of this test? Will the test be designed for or amenable to computer administration? How might differences between versions of the test be reflected in test scores?

- *What is the ideal format of the test?* Should it be true–false, essay, multiple-choice, or in some other format? Why is the format selected for this test the best format?
- *Should more than one form of the test be developed?* On the basis of a cost–benefit analysis, should alternate or parallel forms of this test be created?
- *What special training will be required of test users for administering or interpreting the test?* What background and qualifications will a prospective user of data derived from an administration of this test need to have? What restrictions, if any, should be placed on distributors of the test and on the test's usage?
- *What types of responses will be required of testtakers?* What kind of disability might preclude someone from being able to take this test? What adaptations or accommodations are recommended for persons with disabilities?
- *Who benefits from an administration of this test?* What would the testtaker learn, or how might the testtaker benefit, from an administration of this test? What would the test user learn, or how might the test user benefit? What social benefit, if any, derives from an administration of this test?
- *Is there any potential for harm as the result of an administration of this test?* What safeguards are built into the recommended testing procedure to prevent any sort of harm to any of the parties involved in the use of this test?
- *How will meaning be attributed to scores on this test?* Will a testtaker's score be compared to those of others taking the test at the same time? To those of others in a criterion group? Will the test evaluate mastery of a particular content area?

This last question provides a point of departure for elaborating on issues related to test development with regard to norm- versus criterion-referenced tests.

Norm-referenced versus criterion-referenced tests: Item development issues Different approaches to test development and individual item analyses are necessary, depending upon whether the finished test is designed to be norm-referenced or criterion-referenced. Generally speaking, for example, a good item on a norm-referenced achievement test is an item for which high scorers on the test respond correctly. Low scorers on the test tend to respond to that same item incorrectly. On a criterion-oriented test, this same pattern of results may occur: High scorers on the test get a particular item right whereas low scorers on the test get that same item wrong. However, that is not what makes an item good or acceptable from a criterion-oriented perspective. Ideally, each item on a criterion-oriented test addresses the issue of whether the testtaker—a would-be physician, engineer, piano student, or whoever—has met certain criteria. In short, when it comes to criterion-oriented assessment, being "first in the class" does not count and is often irrelevant. Although we can envision exceptions to this general rule, norm-referenced comparisons typically are insufficient and inappropriate when knowledge of mastery is what the test user requires.

Criterion-referenced testing and assessment are commonly employed in licensing contexts, be it a license to practice medicine or to drive a car. Criterion-referenced approaches are also employed in educational contexts in which mastery of particular material must be demonstrated before the student moves on to advanced material that conceptually builds on the existing base of knowledge, skills, or both.

In contrast to techniques and principles applicable to the development of norm-referenced tests (many of which are discussed in this chapter), the development of criterion-referenced instruments derives from a conceptualization of the knowledge or skills to be mastered. For purposes of assessment, the required cognitive or motor skills may be broken down into

component parts. The test developer may attempt to sample criterion-related knowledge with regard to general principles relevant to the criterion being assessed. Experimentation with different items, tests, formats, or measurement procedures will help the test developer discover the best measure of mastery for the targeted skills or knowledge.

JUST THINK . . .

Suppose you were charged with developing a criterion-referenced test to measure mastery of Chapter 8 of this book. Explain, in as much detail as you think sufficient, how you would go about doing that. It's OK to read on before answering (in fact, you are encouraged to do so).

In general, the development of a criterion-referenced test or assessment procedure may entail exploratory work with at least two groups of testtakers: one group known to have mastered the knowledge or skill being measured and another group known *not* to have mastered such knowledge or skill. For example, during the development of a criterion-referenced written test for a driver's license, a preliminary version of the test may be administered to one group of people who have been driving about 15,000 miles per year for 10 years and who have perfect safety records (no accidents and no moving violations). The second group of testtakers might be a group of adults matched in demographic and related respects to the first group but who have never had any instruction in driving or driving experience. The items that best discriminate between these two groups would be considered "good" items. The preliminary exploratory experimentation done in test development need not have anything at all to do with flying, but you wouldn't know that from its name . . .

Pilot Work

In the context of test development, terms such as **pilot work**, *pilot study,* and *pilot research* refer, in general, to the preliminary research surrounding the creation of a prototype of the test. Test items may be pilot studied (or piloted) to evaluate whether they should be included in the final form of the instrument. In developing a structured interview to measure introversion/extraversion, for example, pilot research may involve open-ended interviews with research subjects believed for some reason (perhaps on the basis of an existing test) to be introverted or extraverted. Additionally, interviews with parents, teachers, friends, and others who know the subject might also be arranged. Another type of pilot study might involve physiological monitoring of the subjects (such as monitoring of heart rate) as a function of exposure to different types of stimuli.

In pilot work, the test developer typically attempts to determine how best to measure a targeted construct. The process may entail literature reviews and experimentation as well as the creation, revision, and deletion of preliminary test items. After pilot work comes the process of test construction. Keep in mind, however, that depending on the nature of the test, as well as the nature of the changing responses to it by testtakers, test users, and the community at large, the need for further pilot research and test revision is always a possibility.

Pilot work is a necessity when constructing tests or other measuring instruments for publication and wide distribution. Of course, pilot work need not be part of the process of developing teacher-made tests for classroom use. Let's take a moment at this juncture to discuss selected aspects of the process of developing tests not for use on the world stage, but rather to measure achievement in a class.

Test Construction

Scaling

We have previously defined *measurement* as the assignment of numbers according to rules. **Scaling** may be defined as the process of setting rules for assigning numbers in measurement. Stated another way, scaling is the process by which a measuring device is designed and

calibrated and by which numbers (or other indices)—scale values—are assigned to different amounts of the trait, attribute, or characteristic being measured.

Historically, the prolific L. L. Thurstone (Figure 8–2) is credited for being at the forefront of efforts to develop methodologically sound scaling methods. He adapted psychophysical scaling methods to the study of psychological variables such as attitudes and values (Thurstone, 1959; Thurstone & Chave, 1929). Thurstone's (1925) article entitled "A Method of Scaling Psychological and Educational Tests" introduced, among other things, the notion of absolute scaling—a procedure for obtaining a measure of item difficulty across samples of testtakers who vary in ability.

Types of scales In common parlance, scales are instruments used to measure something, such as weight. In psychometrics, scales may also be conceived of as instruments used to measure. Here, however, that *something* being measured is likely to be a trait, a state, or an ability. When we think of types of scales, we think of the different ways that scales can be categorized. In Chapter 3, for example, we saw that scales can be meaningfully categorized along a continuum of level of measurement and be referred to as nominal, ordinal, interval, or ratio. But we might also characterize scales in other ways.

If the testtaker's test performance as a function of age is of critical interest, then the test might be referred to as an *age-based scale.* If the testtaker's test performance as a function of grade is of critical interest, then the test might be referred to as a *grade-based scale.* If all raw scores on the test are to be transformed into scores that can range from 1 to 9, then the test might be referred to as a *stanine* scale. A scale might be described in still other ways. For example, it may be categorized as *unidimensional* as opposed to *multidimensional.* It may be categorized as *comparative* as opposed to *categorical.* This is just a sampling of the various ways in which scales can be categorized.

Given that scales can be categorized in many different ways, it would be reasonable to assume that there are many different methods of scaling. Indeed, there are; there is no one method of scaling. There is no best type of scale. Test developers scale a test in the manner they believe is optimally suited to their conception of the measurement of the trait (or whatever) that is being measured.

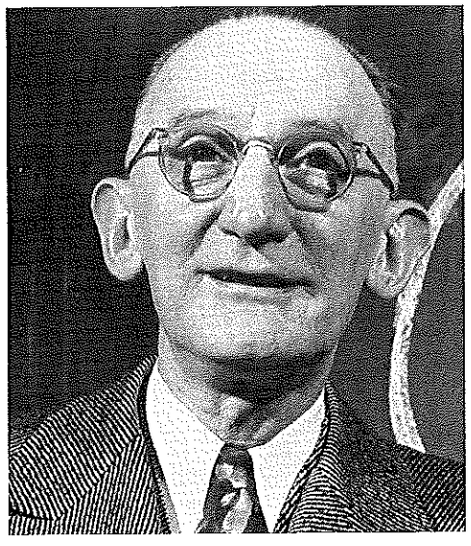

Figure 8–2
L. L. Thurstone (1887–1955)

Among his many achievements in the area of scaling was Thurstone's (1927) influential article "A Law of Comparative Judgment." One of the few "laws" in psychology, this was Thurstone's proudest achievement (Nunnally, 1978, pp. 60–61). Of course, he had many achievements from which to choose. Thurstone's adaptations of scaling methods for use in psychophysiological research and the study of attitudes and values have served as models for generations of researchers (Bock & Jones, 1968). He is also widely considered to be one of the primary architects of modern factor analysis.

Scaling methods Generally speaking, a testtaker is presumed to have more or less of the characteristic measured by a (valid) test as a function of the test score. The higher or lower the score, the more or less of the characteristic the testtaker presumably possesses. But how are numbers assigned to responses so that a test score can be calculated? This is done through scaling the test items, using any one of several available methods.

For example, consider a moral-issues opinion measure called the Morally Debatable Behaviors Scale–Revised (MDBS-R; Katz et al., 1994). Developed to be "a practical means of assessing what people believe, the strength of their convictions, as well as individual differences in moral tolerance" (p. 15), the MDBS-R contains 30 items. Each item contains a brief description of a moral issue or behavior on which testtakers express their opinion by means of a 10-point scale that ranges from "never justified" to "always justified." Here is a sample.

Cheating on taxes if you have a chance is:

1	2	3	4	5	6	7	8	9	10
never justified									always justified

The MDBS-R is an example of a **rating scale,** which can be defined as a grouping of words, statements, or symbols on which judgments of the strength of a particular trait, attitude, or emotion are indicated by the testtaker. Rating scales can be used to record judgments of oneself, others, experiences, or objects, and they can take several forms (Figure 8–3).

On the MDBS-R, the ratings that the testtaker makes for each of the 30 test items are added together to obtain a final score. Scores range from a low of 30 (if the testtaker indicates that all 30 behaviors are never justified) to a high of 300 (if the testtaker indicates that all

Rating Scale Item A
How did you feel about what you saw on television?

Rating Scale Item B
I believe I would like the work of a lighthouse keeper.
True False (circle one)

Rating Scale Item C
Please rate the employee on ability to cooperate and get along with fellow employees:
Excellent _____ /_____ /_____ /_____ /_____ /_____ /_____ / Unsatisfactory

Figure 8–3
The Many Faces of Rating Scales

Rating scales can take many forms. "Smiley" faces, such as those illustrated here as Item A, have been used in social-psychological research with young children and adults with limited language skills. The faces are used in lieu of words such as positive, neutral, and negative.

30 situations are always justified). Because the final test score is obtained by summing the ratings across all the items, it is termed a **summative scale.**

One type of summative rating scale, the **Likert scale** (Likert, 1932), is used extensively in psychology, usually to scale attitudes. Likert scales are relatively easy to construct. Each item presents the testtaker with five alternative responses (sometimes seven), usually on an agree–disagree or approve–disapprove continuum. If Katz et al. had used a Likert scale, an item on their test might have looked like this:

Cheating on taxes if you have a chance.

This is (check one):

never justified	rarely justified	sometimes justified	usually justified	always justified

Likert scales are usually reliable, which may account for their widespread popularity. Likert (1932) experimented with different weightings of the five categories but concluded that assigning weights of 1 (for endorsement of items at one extreme) through 5 (for endorsement of items at the other extreme) generally worked best.

The use of rating scales of any type results in ordinal-level data. With reference to the Likert scale item, for example, if the response *never justified* is assigned the value 1, *rarely justified* the value 2, and so on, then a higher score indicates greater permissiveness with regard to cheating on taxes. Respondents could even be ranked with regard to such permissiveness. However, the difference in permissiveness between the opinions of a pair of people who scored 2 and 3 on this scale is not necessarily the same as the difference between the opinions of a pair of people who scored 3 and 4.

JUST THINK . . .

In your opinion, which version of the Morally Debatable Behaviors Scale is optimal?

Rating scales differ in the number of dimensions underlying the ratings being made. Some rating scales are *unidimensional,* meaning that only one dimension is presumed to underlie the ratings. Other rating scales are *multidimensional,* meaning that more than one dimension is thought to guide the testtaker's responses. Consider in this context an item from the MDBS-R regarding marijuana use. Responses to this item, particularly responses in the low to middle range, may be interpreted in many different ways. Such responses may reflect the view (a) that people should not engage in illegal activities, (b) that people should not take risks with their health, or (c) that people should avoid activities that could lead to contact with a bad crowd. Responses to this item may also reflect other attitudes and beliefs, including those related to documented benefits of marijuana use, as well as new legislation and regulations. When more than one dimension is tapped by an item, multidimensional scaling techniques are used to identify the dimensions.

Another scaling method that produces ordinal data is the **method of paired comparisons.** Testtakers are presented with pairs of stimuli (two photographs, two objects, two statements), which they are asked to compare. They must select one of the stimuli according to some rule; for example, the rule that they agree more with one statement than the other, or the rule that they find one stimulus more appealing than the other. Had Katz et al. used the method of paired comparisons, an item on their scale might have looked like the one that follows.

Select the behavior that you think would be more justified:

a. cheating on taxes if one has a chance

b. accepting a bribe in the course of one's duties

For each pair of options, testtakers receive a higher score for selecting the option deemed more justifiable by the majority of a group of judges. The judges would have been asked to rate the pairs of options before the distribution of the test, and a list of the options selected by the judges would be provided along with the scoring instructions as an answer key. The test score would reflect the number of times the choices of a testtaker agreed with those of the judges. If we use Katz et al.'s (1994) standardization sample as the judges, then the more justifiable option is cheating on taxes. A testtaker might receive a point toward the total score for selecting option "a" but no points for selecting option "b." An advantage of the method of paired comparisons is that it forces testtakers to choose between items.

JUST THINK . . .

Under what circumstance might it be advantageous for tests to contain items presented as a sorting task?

Sorting tasks are another way that ordinal information may be developed and scaled. Here, stimuli such as printed cards, drawings, photographs, or other objects are typically presented to testtakers for evaluation. One method of sorting, **comparative scaling,** entails judgments of a stimulus in comparison with every other stimulus on the scale. A version of the MDBS-R that employs comparative scaling might feature 30 items, each printed on a separate index card. Testtakers would be asked to sort the cards from most justifiable to least justifiable. Comparative scaling could also be accomplished by providing testtakers with a list of 30 items on a sheet of paper and asking them to rank the justifiability of the items from 1 to 30.

Another scaling system that relies on sorting is **categorical scaling.** Stimuli are placed into one of two or more alternative categories that differ quantitatively with respect to some continuum. In our running MDBS-R example, testtakers might be given 30 index cards, on each of which is printed one of the 30 items. Testtakers would be asked to sort the cards into three piles: those behaviors that are never justified, those that are sometimes justified, and those that are always justified.

A **Guttman scale** (Guttman, 1944a,b, 1947) is yet another scaling method that yields ordinal-level measures. Items on it range sequentially from weaker to stronger expressions of the attitude, belief, or feeling being measured. A feature of Guttman scales is that all respondents who agree with the stronger statements of the attitude will also agree with milder statements. Using the MDBS-R scale as an example, consider the following statements that reflect attitudes toward suicide.

Do you agree or disagree with each of the following:

a. All people should have the right to decide whether they wish to end their lives.

b. People who are terminally ill and in pain should have the option to have a doctor assist them in ending their lives.

c. People should have the option to sign away the use of artificial life-support equipment before they become seriously ill.

d. People have the right to a comfortable life.

If this were a perfect Guttman scale, then all respondents who agree with "a" (the most extreme position) should also agree with "b," "c," and "d." All respondents who disagree with "a" but agree with "b" should also agree with "c" and "d," and so forth. Guttman scales are developed through the administration of a number of items to a target group. The resulting data are then analyzed by means of **scalogram analysis**, an item-analysis procedure and approach to test development that involves a graphic mapping of a testtaker's responses. The objective for the developer of a measure of attitudes is to obtain an arrangement of items wherein endorsement of one item automatically connotes endorsement of less extreme positions. It is not always possible to do this. Beyond the measurement of attitudes, Guttman scaling or scalogram analysis (the two terms are used synonymously) appeals to test developers in consumer psychology, where an objective may be to learn if a consumer who will purchase one product will purchase another product.

All the foregoing methods yield ordinal data. The method of equal-appearing intervals, first described by Thurstone (1929), is one scaling method used to obtain data that are presumed to be interval in nature. Again using the example of attitudes about the justifiability of suicide, let's outline the steps that would be involved in creating a scale using Thurstone's equal-appearing intervals method.

1. A reasonably large number of statements reflecting positive and negative attitudes toward suicide are collected, such as *Life is sacred, so people should never take their own lives* and *A person in a great deal of physical or emotional pain may rationally decide that suicide is the best available option.*
2. Judges (or experts in some cases) evaluate each statement in terms of how strongly it indicates that suicide is justified. Each judge is instructed to rate each statement on a scale as if the scale were interval in nature. For example, the scale might range from 1 (the statement indicates that suicide is never justified) to 9 (the statement indicates that suicide is always justified). Judges are instructed that the 1-to-9 scale is being used as if there were an equal distance between each of the values—that is, as if it were an interval scale. Judges are cautioned to focus their ratings on the statements, not on their own views on the matter.
3. A mean and a standard deviation of the judges' ratings are calculated for each statement. For example, if fifteen judges rated 100 statements on a scale from 1 to 9 then, for each of these 100 statements, the fifteen judges' ratings would be averaged. Suppose five of the judges rated a particular item as a 1, five other judges rated it as a 2, and the remaining five judges rated it as a 3. The average rating would be 2 (with a standard deviation of 0.816).
4. Items are selected for inclusion in the final scale based on several criteria, including (a) the degree to which the item contributes to a comprehensive measurement of the variable in question and (b) the test developer's degree of confidence that the items have indeed been sorted into equal intervals. Item means and standard deviations are also considered. Items should represent a wide range of attitudes reflected in a variety of ways. A low standard deviation is indicative of a good item; the judges agreed about the meaning of the item with respect to its reflection of attitudes toward suicide.
5. The scale is now ready for administration. The way the scale is used depends on the objectives of the test situation. Typically, respondents are asked to select those statements that most accurately reflect their own attitudes. The values of the items that the respondent selects (based on the judges' ratings) are averaged, producing a score on the test.

The method of equal-appearing intervals is an example of a scaling method of the *direct estimation* variety. In contrast to other methods that involve *indirect estimation,* there is no need to transform the testtaker's responses into some other scale.

The particular scaling method employed in the development of a new test depends on many factors, including the variables being measured, the group for whom the test is intended (children may require a less complicated scaling method than adults, for example), and the preferences of the test developer.

Writing Items

In the grand scheme of test construction, considerations related to the actual writing of the test's items go hand in hand with scaling considerations. The prospective test developer or item writer immediately faces three questions related to the test blueprint:

- What range of content should the items cover?
- Which of the many different types of item formats should be employed?
- How many items should be written in total and for each content area covered?

When devising a standardized test using a multiple-choice format, it is usually advisable that the first draft contain approximately twice the number of items that the final version of the test will contain.[1] If, for example, a test called "American History: 1940 to 1990" is to have 30 questions in its final version, it would be useful to have as many as 60 items in the item pool. Ideally, these items will adequately sample the domain of the test. An **item pool** is the reservoir or well from which items will or will not be drawn for the final version of the test.

A comprehensive sampling provides a basis for content validity of the final version of the test. Because approximately half of these items will be eliminated from the test's final version, the test developer needs to ensure that the final version also contains items that adequately sample the domain. Thus, if all the questions about the Persian Gulf War from the original 60 items were determined to be poorly written, then the test developer should either rewrite items sampling this period or create new items. The new or rewritten items would then also be subjected to tryout so as not to jeopardize the test's content validity. As in earlier versions of the test, an effort is made to ensure adequate sampling of the domain in the final version of the test. Another consideration here is whether or not alternate forms of the test will be created and, if so, how many. Multiply the number of items required in the pool for one form of the test by the number of forms planned, and you have the total number of items needed for the initial item pool.

How does one develop items for the item pool? The test developer may write a large number of items from personal experience or academic acquaintance with the subject matter. Help may also be sought from others, including experts. For psychological tests designed to be used in clinical settings, clinicians, patients, patients' family members, clinical staff, and others may be interviewed for insights that could assist in item writing. For psychological tests designed to be used by personnel psychologists, interviews with members of a targeted industry or organization will likely be of great value. For psychological tests designed to be used by school psychologists, interviews with teachers, administrative staff, educational psychologists, and others may be invaluable. Searches through the academic research literature may prove fruitful, as may searches through other databases.

Considerations related to variables such as the purpose of the test and the number of examinees to be tested at one time enter into decisions regarding the format of the test under construction.

JUST THINK . . .

If you were going to develop a pool of items to cover the subject of "academic knowledge of what it takes to develop an item pool," how would you go about doing it?

Item format Variables such as the form, plan, structure, arrangement, and layout of individual test items are collectively referred to as **item format.** Two types of item format we will discuss in detail are the *selected-response format* and the *constructed-response format.* Items presented in a **selected-response format** require testtakers to select a response from a set of alternative responses. Items presented in a **constructed-response format** require testtakers to supply or to create the correct answer, not merely to select it.

If a test is designed to measure achievement and if the items are written in a selected-response format, then examinees must select the response that is keyed as correct. If the test is designed to measure the strength of a particular trait and if the items are written in a selected-response format, then examinees must select the alternative that best answers the question with respect to themselves. As we further discuss item formats, for the sake of simplicity we will confine our examples to achievement tests. The reader may wish to mentally substitute other appropriate terms for words such as *correct* for personality or other types of tests that are not achievement tests.

1. Common sense and the practical demands of the situation may dictate that fewer items be written for the first draft of a test. If, for example, the final draft were to contain 1,000 items, then creating an item pool of 2,000 items might be an undue burden. If the test developer is a knowledgeable and capable item writer, it might be necessary to create only about 1,200 items for the item pool.

Three types of selected-response item formats are *multiple-choice*, *matching*, and *true–false*. An item written in a **multiple-choice format** has three elements: (1) a stem, (2) a correct alternative or option, and (3) several incorrect alternatives or options variously referred to as *distractors* or *foils*. Two illustrations follow (despite the fact that you are probably all too familiar with multiple-choice items).

Item A

Stem	⟶	A psychological test, an interview, and a case study are:
Correct alt.	⟶	a. psychological assessment tools
Distractors	⟶	b. standardized behavioral samples c. reliable assessment instruments d. theory-linked measures

Now consider Item B:

Item B

A good multiple-choice item in an achievement test:

a. has one correct alternative
b. has grammatically parallel alternatives
c. has alternatives of similar length
d. has alternatives that fit grammatically with the stem
e. includes as much of the item as possible in the stem to avoid unnecessary repetition
f. avoids ridiculous distractors
g. is not excessively long
h. all of the above
i. none of the above

If you answered "h" to Item B, you are correct. As you read the list of alternatives, it may have occurred to you that Item B violated some of the rules it set forth!

In a **matching item,** the testtaker is presented with two columns: *premises* on the left and *responses* on the right. The testtaker's task is to determine which response is best associated with which premise. For very young testtakers, the instructions will direct them to draw a line from one premise to one response. Testtakers other than young children are typically asked to write a letter or number as a response. Here's an example of a matching item one might see on a test in a class on modern film history:

Directions: Match an actor's name in Column X with a film role the actor played in Column Y. Write the letter of the film role next to the number of the corresponding actor. Each of the roles listed in Column Y may be used once, more than once, or not at all.

Column X	Column Y
_____ 1. Matt Damon	a. Anton Chigurh
_____ 2. Javier Bardem	b. Max Styph
_____ 3. Stephen James	c. Storm
_____ 4. Michael Keaton	d. Jason Bourne
_____ 5. Charlize Theron	e. Ray Kroc
_____ 6. Chris Evans	f. Jesse Owens
_____ 7. George Lazenby	g. Hugh ("The Revenant") Glass
_____ 8. Ben Affleck	h. Steve ("Captain America") Rogers
_____ 9. Keanu Reeves	i. Bruce (Batman) Wayne
_____ 10. Leonardo DiCaprio	j. Aileen Wuornos
_____ 11. Halle Berry	k. James Bond
	l. John Wick
	m. Jennifer Styph

You may have noticed that the two columns contain different numbers of items. If the number of items in the two columns were the same, then a person unsure about one of the actor's roles could merely deduce it by matching all the other options first. A perfect score would then result even though the testtaker did not actually know all the answers. Providing more options than needed minimizes such a possibility. Another way to lessen the probability of chance or guessing as a factor in the test score is to state in the directions that each response may be a correct answer once, more than once, or not at all.

Some guidelines should be observed in writing matching items for classroom use. The wording of the premises and the responses should be fairly short and to the point. No more than a dozen or so premises should be included; otherwise, some students will forget what they were looking for as they go through the lists. The lists of premises and responses should both be homogeneous—that is, lists of the same sort of thing. Our film school example provides a homogeneous list of premises (all names of actors) and a homogeneous list of responses (all names of film characters). Care must be taken to ensure that one and only one premise is matched to one and only one response. For example, adding the name of actors Sean Connery, Roger Moore, David Niven, Timothy Dalton, Pierce Brosnan, or Daniel Craig to the premise column as it now exists would be inadvisable, regardless of what character's name was added to the response column. Do you know why?

At one time or another, Connery, Moore, Niven, Dalton, Brosnan, and Craig all played the role of James Bond (response "k"). As the list of premises and responses currently stands, the match to response "k" is premise "7" (this Australian actor played Agent 007 in the film *On Her Majesty's Secret Service*). If in the future the test developer wanted to substitute the name of another actor—say, Daniel Craig for George Lazenby—then it would be prudent to review the columns to confirm that Craig did not play any of the other characters in the response list and that James Bond still was not played by any actor in the premise list besides Craig.[2]

A multiple-choice item that contains only two possible responses is called a **binary-choice item.** Perhaps the most familiar binary-choice item is the **true–false item.** As you know, this type of selected-response item usually takes the form of a sentence that requires the testtaker to indicate whether the statement is or is not a fact. Other varieties of binary-choice items include sentences to which the testtaker responds with one of two responses, such as *agree or disagree, yes or no, right or wrong,* or *fact or opinion.*

JUST THINK . . .

Respond either true or false, depending upon your opinion as a student: *In the field of education, selected-response items are preferable to constructed-response items.* Then respond again, this time from the perspective of an educator and test user. Explain your answers.

A good binary choice contains a single idea, is not excessively long, and is not subject to debate; the correct response must undoubtedly be one of the two choices. Like multiple-choice items, binary-choice items are readily applicable to a wide range of subjects. Unlike multiple-choice items, binary-choice items cannot contain distractor alternatives. For this reason, binary-choice items are typically easier to write than multiple-choice items and can be written relatively quickly. A disadvantage of the binary-choice item is that the probability of obtaining a correct response purely on the basis of chance (guessing) on any one item is .5, or 50%.[3] In contrast, the probability of obtaining a correct response by guessing on a four-alternative multiple-choice question is .25, or 25%.

2. Here's the entire answer key: 1-d, 2-a, 3-f, 4-e, 5-j, 6-h, 7-k, 8-i, 9-l, 10-g, 11-c.

3. We note in passing, however, that although the probability of guessing correctly on an individual binary-choice item on the basis of chance alone is .5, the probability of guessing correctly on a *sequence* of such items decreases as the number of items increases. The probability of guessing correctly on two such items is equal to $.5^2$, or 25%. The probability of guessing correctly on ten such items is equal to $.5^{10}$, or .001. This means there is a one-in-a-thousand chance that a testtaker would guess correctly on ten true–false (or other binary-choice) items on the basis of chance alone.

Moving from a discussion of the selected-response format to the constructed variety, three types of constructed-response items are the *completion item*, the *short answer*, and the *essay*.

A **completion item** requires the examinee to provide a word or phrase that completes a sentence, as in the following example:

The standard deviation is generally considered the most useful measure of _____.

A good completion item should be worded so that the correct answer is specific. Completion items that can be correctly answered in many ways lead to scoring problems. (The correct completion here is *variability*.) An alternative way of constructing this question would be as a short-answer item:

What descriptive statistic is generally considered the most useful measure of variability?

A completion item may also be referred to as a **short-answer item.** It is desirable for completion or short-answer items to be written clearly enough that the testtaker can respond succinctly—that is, with a short answer. There are no hard-and-fast rules for how short an answer must be to be considered a short answer; a word, a term, a sentence, or a paragraph may qualify. Beyond a paragraph or two, the item is more properly referred to as an essay item. We may define an **essay item** as a test item that requires the testtaker to respond to a question by writing a composition, typically one that demonstrates recall of facts, understanding, analysis, and/or interpretation.

Here is an example of an essay item:

> Compare and contrast definitions and techniques of classical and operant conditioning. Include examples of how principles of each have been applied in clinical as well as educational settings.

An essay item is useful when the test developer wants the examinee to demonstrate a depth of knowledge about a single topic. In contrast to selected-response and constructed-response items such as the short-answer item, the essay question not only permits the restating of learned material but also allows for the creative integration and expression of the material in the testtaker's own words. The skills tapped by essay items are different from those tapped by true–false and matching items. Whereas these latter types of items require only recognition, an essay requires recall, organization, planning, and writing ability. A drawback of the essay item is that it tends to focus on a more limited area than can be covered in the same amount of time when using a series of selected-response items or completion items. Another potential problem with essays can be subjectivity in scoring and inter-scorer differences. A review of some advantages and disadvantages of these different item formats, especially as used in academic classroom settings, is presented in Table 8–1.

Writing items for computer administration A number of widely available computer programs are designed to facilitate the construction of tests as well as their administration, scoring, and interpretation. These programs typically make use of two advantages of digital media: the ability to store items in an *item bank* and the ability to individualize testing through a technique called *item branching*.

An **item bank** is a relatively large and easily accessible collection of test questions. Instructors who regularly teach a particular course sometimes create their own item bank of questions that they have found to be useful on examinations. One of the many potential advantages of an item bank is accessibility to a large number of test items conveniently classified by subject area, item statistics, or other variables. And just as funds may be added to or withdrawn from a more traditional bank, so items may be added to, withdrawn from, and even modified in an item bank. A detailed description of the process of designing an item bank can be found through the Instructor Resources within Connect, in OOBAL-8-B1, "How to 'Fund' an Item Bank."

The term **computerized adaptive testing (CAT)** refers to an interactive, computer-administered test-taking process wherein items presented to the testtaker are based in part on the

Table 8–1
Some Advantages and Disadvantages of Various Item Formats

Format of Item	Advantages	Disadvantages
Multiple-choice	• Can sample a great deal of content in a relatively short time. • Allows for precise interpretation and little "bluffing" other than guessing. This, in turn, may allow for more content-valid test score interpretation than some other formats. • May be machine- or computer-scored.	• Does not allow for expression of original or creative thought. • Not all subject matter lends itself to reduction to one and only one answer keyed correct. • May be time-consuming to construct series of good items. • Advantages of this format may be nullified if item is poorly written or if a pattern of correct alternatives is discerned by the testtaker.
Binary-choice items (such as true/false)	• Can sample a great deal of content in a relatively short time. • Test consisting of such items is relatively easy to construct and score. • May be machine- or computer-scored.	• Susceptibility to guessing is high, especially for "test-wise" students who may detect cues to reject one choice or the other. • Some wordings, including use of adverbs such as *typically* or *usually,* can be interpreted differently by different students. • Can be used only when a choice of dichotomous responses can be made without qualification.
Matching	• Can effectively and efficiently be used to evaluate testtakers' recall of related facts. • Particularly useful when there are a large number of facts on a single topic. • Can be fun or game-like for testtaker (especially the well-prepared testtaker). • May be machine- or computer-scored.	• As with other items in the selected-response format, test-takers need only *recognize* a correct answer and not recall it or devise it. • One of the choices may help eliminate one of the other choices as the correct response. • Requires pools of related information and is of less utility with distinctive ideas.
Completion or short-answer (fill-in-the-blank)	• Wide content area, particularly of questions that require factual recall, can be sampled in relatively brief amount of time. • This type of test is relatively easy to construct. • Useful in obtaining picture of what testtaker is able to generate as opposed to merely recognize since testtaker must generate response.	• Useful only with responses of one word or a few words. • May demonstrate only recall of circumscribed facts or bits of knowledge. • Potential for inter-scorer reliability problems when test is scored by more than one person. • Typically hand-scored.
Essay	• Useful in measuring responses that require complex, imaginative, or original solutions, applications, or demonstrations. • Useful in measuring how well testtaker is able to communicate ideas in writing. • Requires testtaker to generate entire response, not merely recognize it or supply a word or two.	• May not sample wide content area as well as other tests do. • Testtaker with limited knowledge can attempt to bluff with confusing, sometimes long and elaborate writing designed to be as broad and ambiguous as possible. • Scoring can be time-consuming and fraught with pitfalls. • When more than one person is scoring, inter-scorer reliability issues may be raised. • May rely too heavily on writing skills, even to the point of confounding writing ability with what is purportedly being measured. • Typically hand-scored.

JUST THINK . . .

If an item bank is sufficiently large, might it make sense to publish the entire bank of items in advance to the testtakers before the test?

testtaker's performance on previous items. As in traditional test administration, the test might begin with some sample, practice items. However, the computer may not permit the testtaker to continue with the test until the practice items have been responded to in a satisfactory manner and the testtaker has demonstrated an understanding of the test procedure. Using CAT, the test administered may be different for each testtaker, depending on the test performance on the items presented. Each item on an achievement test, for example, may have a known difficulty level. This fact as well as other data (such as a statistical allowance for blind guessing) may be factored in when it comes time to tally a final score on the items administered. Note that we do not say "final score on the test" because what constitutes "the test" may well be different for different testtakers.

The advantages of CAT have been well documented (Weiss & Vale, 1987). Only a sample of the total number of items in the item pool is administered to any one testtaker. On the basis of previous response patterns, items that have a high probability of being answered in a particular fashion ("correctly" if an ability test) are not presented, thus providing economy in terms of testing time and total number of items presented. Computerized adaptive testing has been found to reduce the number of test items that need to be administered by as much as 50% while simultaneously reducing measurement error by 50%.

CAT tends to reduce *floor effects* and *ceiling effects*. A **floor effect** refers to the diminished utility of an assessment tool for distinguishing testtakers at the low end of the ability, trait, or other attribute being measured. A test of ninth-grade mathematics, for example, may contain items that range from easy to hard for testtakers having the mathematical ability of the average ninth-grader. However, testtakers who have not yet achieved such ability might fail all of the items; because of the floor effect, the test would not provide any guidance as to the relative mathematical ability of testtakers in this group. If the item bank contained some less difficult items, these could be pressed into service to minimize the floor effect and provide discrimination among the low-ability testtakers.

As you might expect, a **ceiling effect** refers to the diminished utility of an assessment tool for distinguishing testtakers at the high end of the ability, trait, or other attribute being measured. Returning to our example of the ninth-grade mathematics test, what would happen if all of the testtakers answered all of the items correctly? It is likely that the test user would conclude that the test was too easy for this group of testtakers and so discrimination was impaired by a ceiling effect. If the item bank contained some items that were more difficult, these could be used to minimize the ceiling effect and enable the test user to better discriminate among these high-ability testtakers.

JUST THINK . . .

Provide an example of how a floor effect in a test of integrity might occur when the sample of testtakers consisted of prison inmates convicted of fraud.

The ability of the computer to tailor the content and order of presentation of test items on the basis of responses to previous items is referred to as **item branching.** A computer that has stored a bank of achievement test items of different difficulty levels can be programmed to present items according to an algorithm or rule. For example, one rule might be "don't present an item of the next difficulty level until two consecutive items of the current difficulty level are answered correctly." Another rule might be "terminate the test when five consecutive items of a given level of difficulty have been answered incorrectly." Alternatively, the pattern of items to which the testtaker is exposed might be based not on the testtaker's response to preceding items but on a random drawing from the total pool of test items. Random presentation of items reduces the ease with which testtakers can memorize items on behalf of future testtakers.

JUST THINK . . .

Provide an example of a ceiling effect in a test that measures a personality trait.

Item-branching technology may be applied when constructing tests not only of achievement but also of personality. For example, if a respondent answers an item in a way that suggests he or she is depressed, the computer might automatically probe for depression-related symptoms and behavior. The next item presented might be designed to probe the respondents' sleep patterns or the existence of suicidal ideation.

Item-branching technology may be used in personality tests to recognize nonpurposive or inconsistent responding. For example, on a computer-based true–false test, if the examinee responds *true* to an item such as "I summered in Baghdad last year," then there would be reason to suspect that the examinee is responding nonpurposively, randomly, or in some way other

JUST THINK . . .

Try your hand at writing a couple of true–false items that could be used to detect nonpurposive or random responding on a personality test.

than genuinely. And if the same respondent responds *false* to the identical item later on in the test, the respondent is being inconsistent as well. Should the computer recognize a nonpurposive response pattern, it may be programmed to respond in a prescribed way—for example, by admonishing the respondent to be more careful or even by refusing to proceed until a purposive response is given.

Scoring Items

Many different test scoring models have been devised. Perhaps the model used most commonly—owing, in part, to its simplicity and logic—is the cumulative model. Typically, the rule in a cumulatively scored test is that the higher the score on the test, the higher the testtaker is on the ability, trait, or other characteristic that the test purports to measure. For each testtaker response to targeted items made in a particular way, the testtaker earns cumulative credit with regard to a particular construct.

In tests that employ **class scoring** or (also referred to as **category scoring),** testtaker responses earn credit toward placement in a particular class or category with other testtakers whose pattern of responses is presumably similar in some way. This approach is used by some diagnostic systems wherein individuals must exhibit a certain number of symptoms to qualify for a specific diagnosis. A third scoring model, *ipsative scoring,* departs radically in rationale from either cumulative or class models. A typical objective in **ipsative scoring** is comparing a testtaker's score on one scale within a test to another scale within that same test.

Consider, for example, a personality test called the Edwards Personal Preference Schedule (EPPS), which is designed to measure the relative strength of different psychological needs. The EPPS ipsative scoring system yields information on the strength of various needs in relation to the strength of other needs of the testtaker. The test does not yield information on the strength of a testtaker's need relative to the presumed strength of that need in the general population. Edwards constructed his test of 210 pairs of statements in a way such that respondents were "forced" to answer *true* or *false* or *yes* or *no* to only one of two statements. Prior research by Edwards had indicated that the two statements were equivalent in terms of how socially desirable the responses were. Here is a sample of an EPPS-like forced-choice item, to which the respondents would indicate which is "more true" of themselves:

I feel depressed when I fail at something.

I feel nervous when giving a talk before a group.

On the basis of such an ipsatively scored personality test, it would be possible to draw only intra-individual conclusions about the testtaker. Here's an example: "John's need for achievement is higher than his need for affiliation." It would not be appropriate to draw inter-individual comparisons on the basis of an ipsatively scored test. It would be inappropriate, for example, to compare two testtakers with a statement like "John's need for achievement is higher than Jane's need for achievement."

Once the test developer has decided on a scoring model and has done everything else necessary to prepare the first draft of the test for administration, the next step is test tryout.

Test Tryout

Having created a pool of items from which the final version of the test will be developed, the test developer will try out the test. The test should be tried out on people who are similar in critical respects to the people for whom the test was designed. Thus, for example, if a test is

designed to aid in decisions regarding the selection of corporate employees with management potential at a certain level, it would be appropriate to try out the test on corporate employees at the targeted level.

Equally important are questions about the number of people on whom the test should be tried out. An informal rule of thumb is that there should be no fewer than 5 subjects and preferably as many as 10 for each item on the test. In general, the more subjects in the tryout the better. The thinking here is that the more subjects employed, the weaker the role of chance in subsequent data analysis. A definite risk in using too few subjects during test tryout comes during factor analysis of the findings, when what we might call phantom factors—factors that actually are just artifacts of the small sample size—may emerge.

The test tryout should be executed under conditions as identical as possible to the conditions under which the standardized test will be administered; all instructions, and everything from the time limits allotted for completing the test to the atmosphere at the test site, should be as similar as possible. As Nunnally (1978, p. 279) so aptly phrased it, "If items for a personality inventory are being administered in an atmosphere that encourages frankness and the eventual test is to be administered in an atmosphere where subjects will be reluctant to say bad things about themselves, the item analysis will tell a faulty story." In general, the test developer endeavors to ensure that differences in response to the test's items are due in fact to the items, not to extraneous factors.

JUST THINK . . .

How appropriate would it be to try out a "management potential" test on a convenience sample of introductory psychology students?

In Chapter 4, we dealt in detail with the important question "What is a good test?" Now is a good time to raise a related question.

What Is a Good Item?

Pseudobulbar affect (PBA) is a neurological disorder characterized by frequent and involuntary outbursts of laughing or crying that may or may not be appropriate to the situation. In one study of veterans with traumatic brain injury, the researchers asked whether the respondents had ever experienced exaggerated episodes of laughing or crying. The subjects' responses to this single item were critically important in identifying persons who required more thorough clinical evaluation for PBA symptoms (Rudolph et al., 2016). By any measure, this single survey item about exaggerated laughing or crying constituted, for the purposes of the evaluation, "a good item."

In the same sense that a good test is reliable and valid, a good test item is reliable and valid. Further, a good test item helps to discriminate testtakers. That is, a good test item is one that is answered correctly (or in an expected manner) by high scorers on the test as a whole. Certainly in the context of academic achievement testing, an item that is answered incorrectly by high scorers on the test as a whole is probably not a good item. Conversely, a good test item is one that is answered incorrectly by low scorers on the test as a whole. By the way, it is also the case that an item that is answered correctly by low scorers on the test as a whole may not be a good item.

How does a test developer identify good items? After the first draft of the test has been administered to a representative group of examinees, the test developer analyzes test scores and responses to individual items. The different types of statistical scrutiny that the test data can potentially undergo at this point are referred to collectively as **item analysis.** Although item analysis tends to be regarded as a quantitative endeavor, it may also be qualitative, as we shall see.

JUST THINK . . .

Well, do a bit more than think: Write one good item in any format, along with a brief explanation of why you think it is a good item. The item should be for a new test you are developing called the American History Test, which will be administered to ninth-graders.

Item Analysis

Statistical procedures used to analyze items may become quite complex, and our treatment of this subject should be viewed as only introductory. We briefly survey some procedures typically used by test developers in their efforts to select the best items from a pool of tryout items. The criteria for the best items may differ as a function of the test developer's objectives. Thus, for example, one test developer might deem the best items to be those that optimally contribute to the internal reliability of the test. Another test developer might wish to design a test with the highest possible criterion-related validity and then select items accordingly. Among the tools test developers might employ to analyze and select items are

- an index of the item's difficulty
- an index of the item's reliability
- an index of the item's validity
- an index of item discrimination

JUST THINK . . .

Apply these item-analysis statistics to a test of personality. Make translations in phraseology as you think about how statistics such as an item-difficulty index or an item-validity index could be used to help identify good items for a personality test (not for an achievement test).

Assume for the moment that you got carried away on the previous *Just Think* exercise and are now the proud author of 100 items for a ninth-grade-level American History Test (AHT). Let's further assume that this 100-item (draft) test has been administered to 100 ninth-graders. Hoping in the long run to standardize the test and have it distributed by a commercial test publisher, you have a more immediate, short-term goal: to select the 50 best of the 100 items you originally created. How might that short-term goal be achieved? As we will see, the answer lies in item-analysis procedures.

The Item-Difficulty Index

Suppose every examinee answered item 1 of the AHT correctly. Can we say that item 1 is a good item? What if no one answered item 1 correctly? In either case, item 1 is not a good item. If everyone gets the item right then the item is too easy; if everyone gets the item wrong, the item is too difficult. Just as the test as a whole is designed to provide an index of degree of knowledge about American history, so each individual item on the test should be passed (scored as correct) or failed (scored as incorrect) on the basis of testtakers' differential knowledge of American history.[4]

An index of an item's difficulty is obtained by calculating the proportion of the total number of testtakers who answered the item correctly. A lowercase italic "p" (p) is used to denote item difficulty, and a subscript refers to the item number (so p_1 is read "item-difficulty index for item 1"). The value of an item-difficulty index can theoretically range from 0 (if no one got the item right) to 1 (if everyone got the item right). If 50 of the 100 examinees answered item 2 correctly, then the item-difficulty index for this item would be equal to 50 divided by 100, or .5 ($p_2 = .5$). If 75 of the examinees got item 3 right, then p_3 would be equal to .75 and we could say that item 3 was easier than item 2. Note that the larger the item-difficulty index, the easier the item. Because p refers to the percent of people passing an item, the higher the p for an item, the easier the item. The statistic referred to as an **item-difficulty index** in the context of achievement testing may be an **item-endorsement index** in other contexts, such as personality testing. Here, the

4. An exception here may be a **giveaway item.** Such an item might be inserted near the beginning of an achievement test to spur motivation and a positive test-taking attitude and to lessen testtakers' test-related anxiety. In general, however, if an item analysis suggests that a particular item is too easy or too difficult, the item must be either rewritten or discarded.

statistic provides not a measure of the percent of people passing the item but a measure of the percent of people who said yes to, agreed with, or otherwise endorsed the item.

An index of the difficulty of the average test item for a particular test can be calculated by averaging the item-difficulty indices for all the test's items. This is accomplished by summing the item-difficulty indices for all test items and dividing by the total number of items on the test. For maximum discrimination among the abilities of the testtakers, the optimal average item difficulty is approximately .5, with individual items on the test ranging in difficulty from about .3 to .8. Note, however, that the possible effect of guessing must be taken into account when considering items of the selected-response variety. With this type of item, the optimal average item difficulty is usually the midpoint between 1.00 and the chance success proportion, defined as the probability of answering correctly by random guessing. In a true–false item, the probability of guessing correctly on the basis of chance alone is 1/2, or .50. Therefore, the optimal item difficulty is halfway between .50 and 1.00, or .75. In general, the midpoint representing the optimal item difficulty is obtained by summing the chance success proportion and 1.00 and then dividing the sum by 2, or

JUST THINK . . .

Create an achievement test item having to do with any aspect of psychological testing and assessment that you believe would yield a *p* of 0 if administered to every member of your class.

$$.5 + 1.00 = 1.5$$
$$\frac{1.5}{2} = .60$$

For a five-option multiple-choice item, the probability of guessing correctly on any one item on the basis of chance alone is equal to 1/5, or .20. The optimal item difficulty is therefore .60:

$$.20 + 1.00 = 1.20$$
$$\frac{1.20}{2} = .60$$

The Item-Reliability Index

The **item-reliability index** provides an indication of the internal consistency of a test (Figure 8–4); the higher this index, the greater the test's internal consistency. This index is equal to the product of the item-score standard deviation (s) and the correlation (r) between the item score and the total test score.

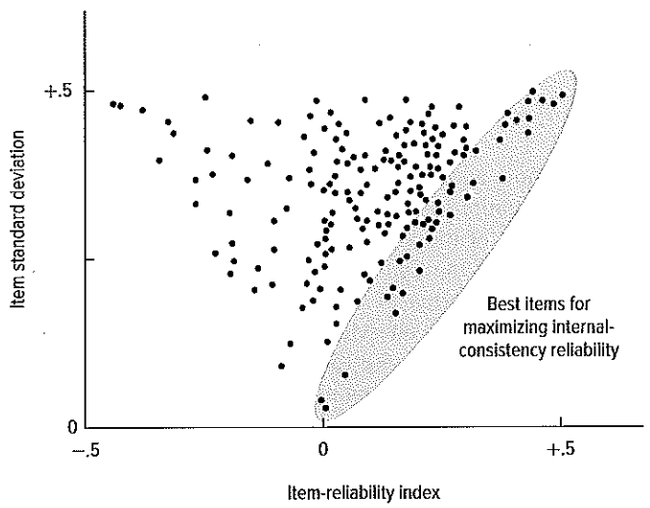

Figure 8–4
Maximizing Internal-Consistency Reliability

Source: Allen and Yen (1979).

JUST THINK . . .

An achievement test on the subject of test development is designed to have two items that load on a factor called "item analysis." Write these two test items.

Factor analysis and inter-item consistency A statistical tool useful in determining whether items on a test appear to be measuring the same thing(s) is factor analysis. Through the judicious use of factor analysis, items that do not "load on" the factor that they were written to tap (or, items that do not appear to be measuring what they were designed to measure) can be revised or eliminated. If too many items appear to be tapping a particular area, the weakest of such items can be eliminated. Additionally, factor analysis can be useful in the test interpretation process, especially when comparing the constellation of responses to the items from two or more groups. Thus, for example, if a particular personality test is administered to two groups of hospitalized psychiatric patients, each group with a different diagnosis, then the same items may be found to load on different factors in the two groups. Such information will compel the responsible test developer to revise or eliminate certain items from the test or to describe the differential findings in the test manual.

The Item-Validity Index

The **item-validity index** is a statistic designed to provide an indication of the degree to which a test is measuring what it purports to measure. The higher the item-validity index, the greater the test's criterion-related validity. The item-validity index can be calculated once the following two statistics are known:

- the item-score standard deviation
- the correlation between the item score and the criterion score

The item-score standard deviation of item 1 (denoted by the symbol s_1) can be calculated using the index of the item's difficulty (p_1) in the following formula:

$$s_1 = \sqrt{p_1(1 - p_1)}$$

The correlation between the score on item 1 and a score on the criterion measure (denoted by the symbol $r_{1\,C}$) is multiplied by item 1's item-score standard deviation (s_1), and the product is equal to an index of an item's validity ($s_1\ r_{1\,C}$). Calculating the item-validity index will be important when the test developer's goal is to maximize the criterion-related validity of the test. A visual representation of the best items on a test (if the objective is to maximize criterion-related validity) can be achieved by plotting each item's item-validity index and item-reliability index (Figure 8–5).

The Item-Discrimination Index

Measures of item discrimination indicate how adequately an item separates or discriminates between high scorers and low scorers on an entire test. In this context, a multiple-choice item on an achievement test is a good item if most of the high scorers answer correctly and most of the low scorers answer incorrectly. If most of the high scorers fail a particular item, these testtakers may be making an alternative interpretation of a response intended to serve as a distractor. In such a case, the test developer should interview the examinees to understand better the basis for the choice and then appropriately revise (or eliminate) the item. Common sense dictates that an item on an achievement test is not doing its job if it is answered correctly by respondents who least understand the subject matter. Similarly, an item on a test purporting to measure a particular personality trait is not doing its job if responses indicate that people who score very low on the test as a whole (indicating absence or low levels of the trait in

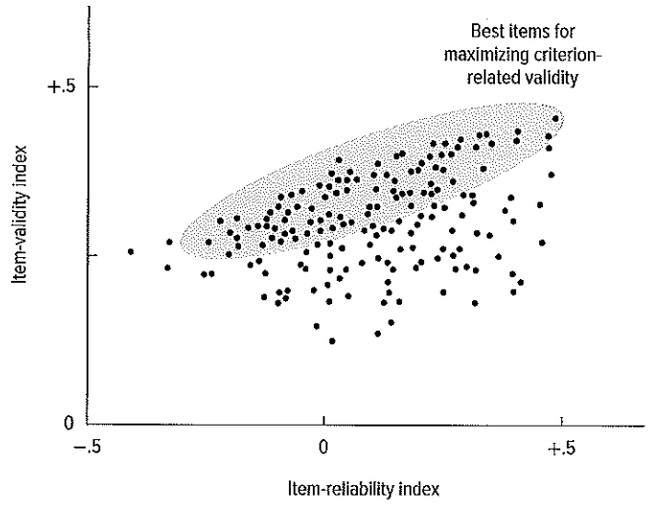

Figure 8–5
Maximizing Criterion-Related Validity

Source: Allen and Yen (1979).

question) tend to score very high on the item (indicating that they are very high on the trait in question—contrary to what the test as a whole indicates).

The **item-discrimination index** is a measure of item discrimination, symbolized by a lowercase italic "d" (*d*). This estimate of item discrimination, in essence, compares performance on a particular item with performance in the upper and lower regions of a distribution of continuous test scores. The optimal boundary lines for what we refer to as the "upper" and "lower" areas of a distribution of scores will demarcate the upper and lower 27% of the distribution of scores—provided the distribution is normal (Kelley, 1939). As the distribution of test scores becomes more platykurtic (flatter), the optimal boundary line for defining upper and lower increases to near 33% (Cureton, 1957). Allen and Yen (1979, p. 122) assure us that "for most applications, any percentage between 25 and 33 will yield similar estimates."

The item-discrimination index is a measure of the difference between the proportion of high scorers answering an item correctly and the proportion of low scorers answering the item correctly; the higher the value of *d*, the greater the number of high scorers answering the item correctly. A negative *d*-value on a particular item is a red flag because it indicates that low-scoring examinees are more likely to answer the item correctly than high-scoring examinees. This situation calls for some action such as revising or eliminating the item.

JUST THINK . . .

Write two items on the subject of test development. The first item to be one that you will predict will have a very high *d*, and the second to be one that you predict will have a high negative *d*.

Suppose a history teacher gave the AHT to a total of 119 students who were just weeks away from completing ninth grade. The teacher isolated the upper (*U*) and lower (*L*) 27% of the test papers, with a total of 32 papers in each group. Data and item-discrimination indices for Items 1 through 5 are presented in Table 8–2. Observe that 20 testtakers in the *U* group answered Item 1 correctly and that 16 testtakers in the *L* group answered Item 1 correctly. With an item-discrimination index equal to .13, Item 1 is probably a reasonable item because more *U*-group members than *L*-group members answered it correctly. The higher the value of *d*, the more adequately the item discriminates the higher-scoring from the lower-scoring testtakers. For this reason, Item 2 is a better item than Item 1 because Item 2's item-discrimination index is .63. The highest possible value

Table 8–2
Item-Discrimination Indices for Five Hypothetical Items

Item	U	L	$U - L$	n	$d[(U - L)/n]$
1	20	16	4	32	.13
2	30	10	20	32	.63
3	32	0	32	32	1.00
4	20	20	0	32	0.00
5	0	32	−32	32	−1.00

of d is +1.00. This value indicates that all members of the U group answered the item correctly whereas all members of the L group answered the item incorrectly.

If the same proportion of members of the U and L groups pass the item, then the item is not discriminating between testtakers at all and d, appropriately enough, will be equal to 0. The lowest value that an index of item discrimination can take is −1. A d equal to −1 is a test developer's nightmare: It indicates that all members of the U group failed the item and all members of the L group passed it. On the face of it, such an item is the worst possible type of item and is in dire need of revision or elimination. However, through further investigation of this unanticipated finding, the test developer might learn or discover something new about the construct being measured.

Analysis of item alternatives The quality of each alternative within a multiple-choice item can be readily assessed with reference to the comparative performance of upper and lower scorers. No formulas or statistics are necessary here. By charting the number of testtakers in the U and L groups who chose each alternative, the test developer can get an idea of the effectiveness of a distractor by means of a simple eyeball test. To illustrate, let's analyze responses to five items on a hypothetical test, assuming that there were 32 scores in the upper level (U) of the distribution and 32 scores in the lower level (L) of the distribution. Let's begin by looking at the pattern of responses to item 1. In each case, ♦ denotes the correct alternative.

		Alternatives				
Item 1		♦a	b	c	d	e
	U	24	3	2	0	3
	L	10	5	6	6	5

The response pattern to Item 1 indicates that the item is a good one. More U group members than L group members answered the item correctly, and each of the distractors attracted some testtakers.

		Alternatives				
Item 2		a	b	c	d	♦e
	U	2	13	3	2	12
	L	6	7	5	7	7

Item 2 signals a situation in which a relatively large number of members of the U group chose a particular distractor choice (in this case, "b"). This item could probably be improved upon revision, preferably one made after an interview with some or all of the U students who chose "b."

Item 3		Alternatives				
		a	b	♦c	d	e
	U	0	0	32	0	0
	L	3	2	22	2	3

Item 3 indicates a most desirable pattern of testtaker response. All members of the *U* group answered the item correctly, and each distractor attracted one or more members of the *L* group.

Item 4		Alternatives				
		a	♦b	c	d	e
	U	5	15	0	5	7
	L	4	5	4	4	14

Item 4 is more difficult than Item 3; fewer examinees answered it correctly. Still, this item provides useful information because it effectively discriminates higher-scoring from lower-scoring examinees. For some reason, one of the alternatives ("e") was particularly effective—perhaps too effective—as a distractor to students in the low-scoring group. The test developer may wish to further explore why this was the case.

Item 5		Alternatives				
		a	b	c	♦d	e
	U	14	0	0	5	13
	L	7	0	0	16	9

Item 5 is a poor item because more *L* group members than *U* group members answered the item correctly. Furthermore, none of the examinees chose the "b" or "c" distractors.

Before moving on to a consideration of the use of item-characteristic curves in item analysis, let's pause to "bring home" the real-life application of some of what we have discussed so far. In his capacity as a consulting industrial/organizational psychologist, our featured test user in this chapter, Dr. Scott Birkeland, has had occasion to create tests and improve them with item-analytic methods. He shares some of his thoughts in his *Meet an Assessment Professional* essay, an excerpt of which is presented here.

Item-Characteristic Curves

As you may have surmised from the introduction to item response theory (IRT) that was presented in Chapter 5, IRT can be a powerful tool not only for understanding how test items perform but also for creating or modifying individual test items, building new tests, and revising existing tests. We will have more to say about that later in the chapter. For now, let's review how *item-characteristic curves (ICCs)* can play a role in decisions about which items are working well and which items are not. Recall that an **item-characteristic curve** is a graphic representation of item difficulty and discrimination.

Figure 8–6 presents several ICCs with ability plotted on the horizontal axis and probability of correct response plotted on the vertical axis. Note that the extent to which an item discriminates high- from low-scoring examinees is apparent from the slope of the curve. The steeper the slope, the greater the item discrimination. An item may also vary in terms of its difficulty level. An easy item will shift the ICC to the left along the ability axis, indicating that many people will likely get the item correct. A difficult item will shift the ICC to the right along the horizontal axis, indicating that fewer people will answer the item correctly. In other words, it takes high ability levels for a person to have a high probability of their response being scored as correct.

Now focus on the item-characteristic curve for Item A. Do you think this is a good item? The answer is that it is not. The probability of a testtaker's responding correctly is high for

MEET AN ASSESSMENT PROFESSIONAL

Meet Dr. Scott Birkeland

I also get involved in developing new test items. Given that these tests are used with real-life candidates, I place a high level of importance on a test's face validity. I want applicants who take the tests to walk away feeling as though the questions that they answered were truly relevant for the job for which they applied. Because of this, each new project leads to the development of new questions so that the tests "look and feel right" for the candidates. For example, if we have a reading and comprehension test, we make sure that the materials that the candidates read are materials that are similar to what they would actually read on the job. This can be a challenge in that by having to develop new questions, the test development process takes more time and effort. In the long run, however, we know that this enhances the candidates' reactions to the testing process. Additionally, our research suggests that it enhances the test's predictability.

Once tests have been developed and administered to candidates, we continue to look for ways to improve them. This is where statistics comes into play. We conduct item level analyses of each question to determine if certain questions are performing better than others. I am often amazed at the power of a simple item analysis (or, calculating item difficulty and item discrimination). Oftentimes, an item analysis will flag a question, causing me to go back and re-examine the item only to find something about it to be confusing. An item analysis allows us to fix those types of issues and continually enhance the quality of a test.

Scott Birkeland, Ph.D., Stang Decision Systems, Inc.

© Scott Birkeland

Read more of what Dr. Birkeland had to say—his complete essay—through the Instructor Resources within Connect.

Used with permission of Scott Birkeland.

testtakers of low ability and low for testtakers of high ability. What about Item B; is it a good test item? Again, the answer is no. The curve tells us that testtakers of moderate ability have the highest probability of answering this item correctly. Testtakers with the greatest amount of ability—as well as their counterparts at the other end of the ability spectrum—are unlikely to respond correctly to this item. Item B may be one of those items to which people who know too much (or think too much) are likely to respond incorrectly.

Item C is a good test item because the probability of responding correctly to it increases with ability. What about Item D? Its ICC profiles an item that discriminates at only one point on the continuum of ability. The probability is great that all testtakers at or above this point will respond correctly to the item, and the probability of an incorrect response is great for testtakers who fall below that particular point in ability. An item such as D therefore has

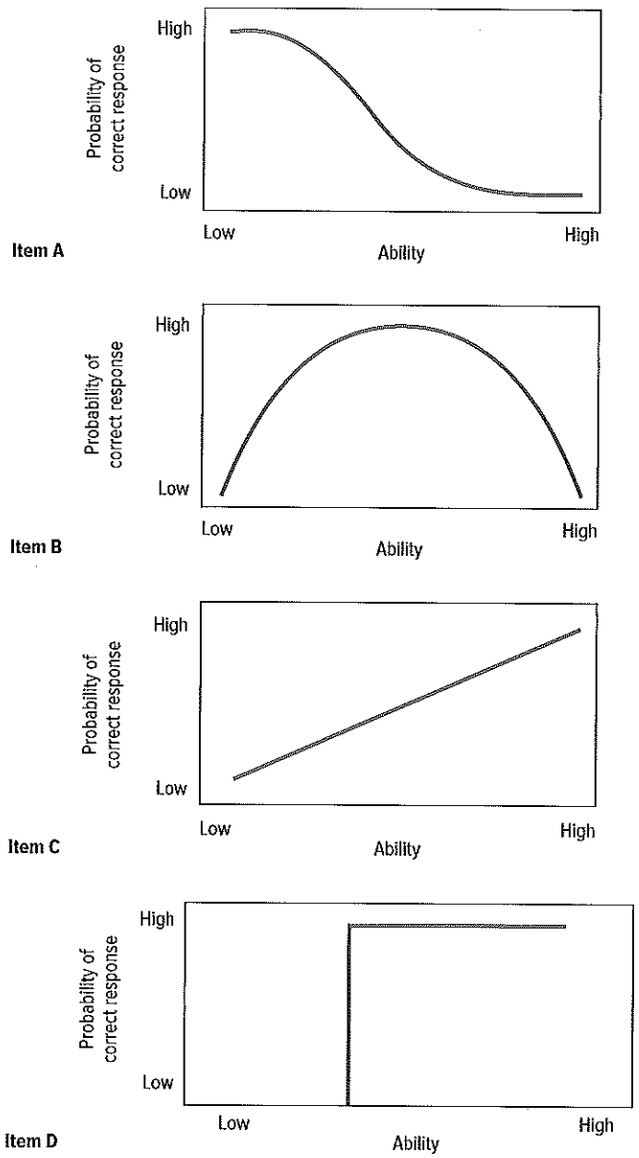

Figure 8–6
Some Sample Item-Characteristic Curves

For simplicity, we have omitted scale values for the axes. The vertical axis in such a graph lists probability of correct response in values ranging from 0 to 1. Values for the horizontal axis, which we have simply labeled "ability," are total scores on the test. In other sources, you may find the vertical axis of an item-characteristic curve labeled something like "proportion of examinees who respond correctly to the item" and the horizontal axis labeled "total test score."

Source: Ghiselli et al. (1981).

excellent discriminative ability and would be useful in a test designed, for example, to select applicants on the basis of some cutoff score. However, such an item might not be desirable in a test designed to provide detailed information on testtaker ability across all ability levels. This might be the case, for example, in a diagnostic reading or arithmetic test.

Other Considerations in Item Analysis

Guessing In achievement testing, the problem of how to handle testtaker **guessing** is one that has eluded any universally acceptable solution. Methods designed to detect guessing (S.-R. Chang et al., 2011), minimize the effects of guessing (Kubinger et al., 2010), and statistically correct for guessing (Espinosa & Gardeazabal, 2010) have been proposed, but no such method has achieved universal acceptance. Perhaps it is because the issues surrounding guessing are more complex than they appear at first glance. To better appreciate the complexity of the issues, consider the following three criteria that any correction for guessing must meet as well as the other interacting issues that must be addressed:

1. A correction for guessing must recognize that, when a respondent guesses at an answer on an achievement test, the guess is not typically made on a totally random basis. It is more reasonable to assume that the testtaker's guess is based on some knowledge of the subject matter and the ability to rule out one or more of the distractor alternatives. However, the individual testtaker's amount of knowledge of the subject matter will vary from one item to the next.
2. A correction for guessing must also deal with the problem of omitted items. Sometimes, instead of guessing, the testtaker will simply omit a response to an item. Should the omitted item be scored "wrong"? Should the omitted item be excluded from the item analysis? Should the omitted item be scored as if the testtaker had made a random guess? Exactly how should the omitted item be handled?
3. Just as some people may be luckier than others in front of a Las Vegas slot machine, so some testtakers may be luckier than others in guessing the choices that are keyed correct. Any correction for guessing may seriously underestimate or overestimate the effects of guessing for lucky and unlucky testtakers.

In addition to proposed interventions at the level of test scoring through the use of corrections for guessing (referred to as formula scores), intervention has also been proposed at the level of test instructions. Testtakers may be instructed to provide an answer only when they are certain (no guessing) or to complete all items and guess when in doubt. Individual differences in testtakers' willingness to take risks result in problems for this approach to guessing (Slakter et al., 1975). Some people who don't mind taking risks may guess even when instructed not to do so. Others who tend to be reluctant to take risks refuse to guess under any circumstances. This creates a situation in which predisposition to take risks can affect one's test score.

JUST THINK . . .

The prevailing logic among measurement professionals is that when testtakers guess at an answer on a personality test in a selected-response format, the testtaker is making the best choice. Why should professionals continue to believe this? Alternatively, why might they modify their view?

To date, no solution to the problem of guessing has been deemed entirely satisfactory. The responsible test developer addresses the problem of guessing by including in the test manual (1) explicit instructions regarding this point for the examiner to convey to the examinees and (2) specific instructions for scoring and interpreting omitted items.

Guessing on responses to personality and related psychological tests is not thought of as a great problem. Although it may sometimes be difficult to choose the most appropriate alternative on a selected-response format personality test (particularly one with forced-choice items), the presumption is that the testtaker does indeed make the best choice.

Item fairness Just as we may speak of biased tests, we may speak of biased test items. The term **item fairness** refers to the degree, if any, a test item is biased. A **biased test item** is an item that favors one particular group of examinees in relation to another when differences in group ability are controlled (Camilli & Shepard, 1985). Many different methods may be used

to identify biased test items. In fact, evidence suggests that the choice of item-analysis method may affect determinations of item bias (Ironson & Subkoviak, 1979).

Item-characteristic curves can be used to identify biased items. Specific items are identified as biased in a statistical sense if they exhibit differential item functioning. Differential item functioning is exemplified by different shapes of item-characteristic curves for different groups (say, men and women) when the two groups do not differ in total test score (Mellenbergh, 1994). If an item is to be considered fair to different groups of testtakers, the item-characteristic curves for the different groups should not be significantly different:

> The essential rationale of this ICC criterion of item bias is that any persons showing the same ability as measured by the whole test should have the same probability of passing any given item that measures that ability, regardless of the person's race, social class, sex, or any other background characteristics. In other words, the same proportion of persons from each group should pass any given item of the test, provided that the persons all earned the same total score on the test. (Jensen, 1980, p. 444)

Establishing the presence of differential item functioning requires a statistical test of the null hypothesis of no difference between the item-characteristic curves of the two groups. The pros and cons of different statistical tests for detecting differential item functioning have long been a matter of debate (Raju et al., 1993). What is not a matter of debate is that items exhibiting significant difference in item-characteristic curves must be revised or eliminated from the test. If a relatively large number of items biased in favor of one group coexist with approximately the same number of items biased in favor of another group, it cannot be claimed that the test measures the same abilities in the two groups. This is true even though overall test scores of the individuals in the two groups may not be significantly different (Jensen, 1980).

JUST THINK . . .

Write an item that is purposely designed to be biased in favor of one group over another. Members of what group would do well on this item? Members of what group would do poorly on this item?

Speed tests Item analyses of tests taken under speed conditions yield misleading or uninterpretable results. The closer an item is to the end of the test, the more difficult it may appear to be. This is because testtakers simply may not get to items near the end of the test before time runs out.

In a similar vein, measures of item discrimination may be artificially high for late-appearing items. This is so because testtakers who know the material better may work faster and are thus more likely to answer the later items. Items appearing late in a speed test are consequently more likely to show positive item-total correlations because of the select group of examinees reaching those items.

Given these problems, how can items on a speed test be analyzed? Perhaps the most obvious solution is to restrict the item analysis of items on a speed test only to the items completed by the testtaker. However, this solution is not recommended, for at least three reasons: (1) Item analyses of the later items would be based on a progressively smaller number of testtakers, yielding progressively less reliable results; (2) if the more knowledgeable examinees reach the later items, then part of the analysis is based on all testtakers and part is based on a selected sample; and (3) because the more knowledgeable testtakers are more likely to score correctly, their performance will make items occurring toward the end of the test appear to be easier than they are.

JUST THINK . . .

Provide an example of what, in your opinion is the best, as well as the worst, use of a speed test.

If speed is not an important element of the ability being measured by the test, and because speed as a variable may produce misleading information about item performance, the test developer ideally should administer the test to be item-analyzed

with generous time limits to complete the test. Once the item analysis is completed, norms should be established using the speed conditions intended for use with the test in actual practice.

Qualitative Item Analysis

Test users have had a long-standing interest in understanding test performance from the perspective of testtakers (Fiske, 1967; Mosier, 1947). The calculation of item-validity, item-reliability, and other such *quantitative* indices represents one approach to understanding testtakers. Another general class of research methods is referred to as *qualitative.* In contrast to quantitative methods, **qualitative methods** are techniques of data generation and analysis that rely primarily on verbal rather than mathematical or statistical procedures. Encouraging testtakers—on a group or individual basis—to discuss aspects of their test-taking experience is, in essence, eliciting or generating "data" (words). These data may then be used by test developers, users, and publishers to improve various aspects of the test.

Qualitative item analysis is a general term for various nonstatistical procedures designed to explore how individual test items work. The analysis compares individual test items to each other and to the test as a whole. In contrast to statistically based procedures, qualitative methods involve exploration of the issues through verbal means such as interviews and group discussions conducted with testtakers and other relevant parties. Some of the topics researchers may wish to explore qualitatively are summarized in Table 8–3.

One cautionary note: Providing testtakers with the opportunity to describe a test can be like providing students with the opportunity to describe their instructors. In both cases, there may be abuse of the process, especially by respondents who have extra-test (or extra-instructor) axes to grind. Respondents may be disgruntled for any number of reasons, from failure to prepare adequately for the test to disappointment in their test performance. In such cases, the opportunity to evaluate the test is an opportunity to lash out. The test, the administrator of the test, and the institution, agency, or corporation responsible for the test administration may all become objects of criticism. Testtaker questionnaires, much like other qualitative research tools, must be interpreted with an eye toward the full context of the experience for the respondent(s).

"Think aloud" test administration An innovative approach to cognitive assessment entails having respondents verbalize thoughts as they occur. Although different researchers use different procedures (Davison et al., 1997; Hurlburt, 1997; Klinger, 1978), this general approach has been employed in a variety of research contexts, including studies of adjustment (Kendall et al., 1979; Sutton-Simon & Goldfried, 1979), problem solving (Duncker, 1945; Kozhevnikov et al., 2007; Montague, 1993), educational research and remediation (Muñoz et al., 2006; Randall et al., 1986; Schellings et al., 2006), clinical intervention (Gann & Davison, 1997; Haaga et al., 1993; Schmitter-Edgecombe & Bales, 2005; White et al., 1992), and jury modeling (Wright & Hall, 2007).

Cohen et al. (1988) proposed the use of **"think aloud" test administration** as a qualitative research tool designed to shed light on the testtaker's thought processes during the administration of a test. On a one-to-one basis with an examiner, examinees are asked to take a test, thinking aloud as they respond to each item. If the test is designed to measure achievement, such verbalizations may be useful in assessing not only if certain students (such as low or high scorers on previous examinations) are misinterpreting a particular item but also *why* and *how* they are misinterpreting the item. If the test is designed to measure personality or some aspect of it, the "think aloud" technique may also yield valuable insights regarding the way individuals perceive, interpret, and respond to the items.

JUST THINK (ALOUD) . . .

How might thinking aloud to evaluate test items be more effective than thinking silently?

Table 8–3
Potential Areas of Exploration by Means of Qualitative Item Analysis

This table lists sample topics and questions of possible interest to test users. The questions could be raised either orally or in writing shortly after a test's administration. Additionally, depending upon the objectives of the test user, the questions could be placed into other formats, such as true–false or multiple choice. Depending upon the specific questions to be asked and the number of testtakers being sampled, the test user may wish to guarantee the anonymity of the respondents.

Topic	Sample Question
Cultural Sensitivity	Did you feel that any item or aspect of this test was discriminatory with respect to any group of people? If so, why?
Face Validity	Did the test appear to measure what you expected it would measure? If not, what was contrary to your expectations?
Test Administrator	Did the behavior of the test administrator affect your performance on this test in any way? If so, how?
Test Environment	Did any conditions in the room affect your performance on this test in any way? If so, how?
Test Fairness	Do you think the test was a fair test of what it sought to measure? Why or why not?
Test Language	Were there any instructions or other written aspects of the test that you had difficulty understanding?
Test Length	How did you feel about the length of the test with respect to (a) the time it took to complete and (b) the number of items?
Testtaker's Guessing	Did you guess on any of the test items? What percentage of the items would you estimate you guessed on? Did you employ any particular strategy for guessing, or was it basically random?
Testtaker's Integrity	Do you think that there was any cheating during this test? If so, please describe the methods you think may have been used.
Testtaker's Mental/Physical State Upon Entry	How would you describe your mental state going into this test? Do you think that your mental state in any way affected the test outcome? If so, how? How would you describe your physical state going into this test? Do you think that your physical state in any way affected the test outcome? If so, how?
Testtaker's Mental/Physical State During the Test	How would you describe your mental state as you took this test? Do you think that your mental state in any way affected the test outcome? If so, how? How would you describe your physical state as you took this test? Do you think that your physical state in any way affected the test outcome? If so, how?
Testtaker's Overall Impressions	What is your overall impression of this test? What suggestions would you offer the test developer for improvement?
Testtaker's Preferences	Did you find any part of the test educational, entertaining, or otherwise rewarding? What, specifically, did you like or dislike about the test? Did you find any part of the test anxiety-provoking, condescending, or otherwise upsetting? Why?
Testtaker's Preparation	How did you prepare for this test? If you were going to advise others how to prepare for it, what would you tell them?

Expert panels In addition to interviewing testtakers individually or in groups, **expert panels** may also provide qualitative analyses of test items. A **sensitivity review** is a study of test items, typically conducted during the test development process, in which items are examined for fairness to all prospective testtakers and for the presence of offensive language, stereotypes, or situations. Since the 1990s or so, sensitivity reviews have become a standard part of test development (Reckase, 1996). For example, in an effort to root out any possible bias in the Stanford Achievement Test series, the test publisher formed an advisory panel of twelve minority group members, each a prominent member of the educational community. Panel

members met with the publisher to obtain an understanding of the history and philosophy of the test battery and to discuss and define the problem of bias. Some of the possible forms of content bias that may find their way into any achievement test were identified as follows (Stanford Special Report, 1992, pp. 3–4).

> *Status:* Are the members of a particular group shown in situations that do not involve authority or leadership?
>
> *Stereotype:* Are the members of a particular group portrayed as uniformly having certain (1) aptitudes, (2) interests, (3) occupations, or (4) personality characteristics?
>
> *Familiarity:* Is there greater opportunity on the part of one group to (1) be acquainted with the vocabulary or (2) experience the situation presented by an item?
>
> *Offensive Choice of Words:* (1) Has a demeaning label been applied, or (2) has a male term been used where a neutral term could be substituted?
>
> *Other:* Panel members were asked to be specific regarding any other indication of bias they detected.

Expert panels may also play a role in the development of new tools of assessment for members of underserved populations. Additionally, experts on a particular culture can inform test developers on optimal ways to achieve desired measurement ends with specific populations of testtakers. This chapter's *Everyday Psychometrics* provides a unique and fascinating glimpse into the process of developing evaluative tools for use with Aboriginal tribe members.

On the basis of qualitative information from an expert panel or testtakers themselves, a test user or developer may elect to modify or revise the test. In this sense, revision typically involves rewording items, deleting items, or creating new items. Note that there is another meaning of test revision beyond that associated with a stage in the development of a new test. After a period of time, many existing tests are scheduled for republication in new versions or editions. The development process that the test undergoes as it is modified and revised is called, not surprisingly, *test revision*. The time, effort, and expense entailed by this latter variety of test revision may be quite extensive. For example, the revision may involve an age extension of the population for which the test is designed for use—upward for older testtakers and/or downward for younger testtakers—and corresponding new validation studies.

JUST THINK . . .

Is there any way that expert panels might introduce more error into the test development process?

Test Revision

We first consider aspects of test revision as a stage in the development of a new test. Later we will consider aspects of test revision in the context of modifying an existing test to create a new edition. Much of our discussion of test revision in the development of a brand-new test may also apply to the development of subsequent editions of existing tests, depending on just how "revised" the revision really is.

Test Revision as a Stage in New Test Development

Having conceptualized the new test, constructed it, tried it out, and item-analyzed it both quantitatively and qualitatively, what remains is to act judiciously on all the information and mold the test into its final form. A tremendous amount of information is generated at the item-analysis stage, particularly given that a developing test may have hundreds of items. On

EVERYDAY PSYCHOMETRICS*

Adapting Tools of Assessment for Use with Specific Cultural Groups

Imagine the cultural misunderstandings that may arise when an assessor with a Western perspective evaluates someone from a non-Western culture. As a case in point, consider the potential for serial misinterpretation of signs and symptoms if the assessor is a Caucasian Westerner and the assesse is a member of an Australian Indigenous culture (commonly referred to in Australia and elsewhere as Aboriginal and Torres Strait Islander people) being evaluated for depression.

For Indigenous Australians, health is viewed in a holistic context—one that encompasses not only mental and physical aspects but cultural and spiritual aspects as well. Ill health is often conceived of as a disruption of these interrelated domains. Perhaps consequently, an Indigenous Australian person is more likely to be perceived in the eyes of a Western evaluator, as presenting with vague complaints of illness—this as opposed to more specific symptomatology. Also, shyness is common in the Indigenous Australian population. Shyness during a mental status examination or other evaluation may manifest itself by avoidance of eye contact with the examiner, which, in turn, may be misinterpreted by the examiner as pathological or otherwise suspect behavior. Another potentially misleading sign or symptom of psychopathology could be the respondent's delayed answers and only minimal speech. However, what might otherwise be interpreted as psychomotor retardation or poverty of speech may well have a cultural basis. Traditional Indigenous Australian people are frequently reserved with, and seemingly indifferent to, Caucasian clinicians, especially in a one-on-one assessment situation. Patients who exhibit a blank or unreactive expression may "come alive" with appropriate affect when a family member or two joins the interview.

Knowledge of Aboriginal culture and clinical experience has suggested to us that when interviewing members of this group, a *yarning* approach works best. Loosely defined, the yarning approach is an interview strategy characterized by the creation of an atmosphere conducive to interviewees conversationally telling their own stories in their own ways. In stark contrast to yarning would be an interview characterized by interrogation, where one direct question is posed after another.

In developing a mental health screening tool for use with members of the Aboriginal culture, a group of clinicians and academic psychiatrists from metropolitan and rural areas of Western Australia and the Northern Territory employed the yarning approach. The interview tool, called the "Here and Now Aboriginal Assessment" (HANAA; see Janca et al., 2015), allows for a traditional story-telling style that involves both family and social yarning. An objective of the design of the instrument was to obtain more meaningful reporting of individual problems while still gathering culturally relevant information about an interviewee's collective identity. **Anhedonia** (inability to experience happiness) may be explored by asking questions such as "Have you lost interest in things that you used to like doing?" Engagement in culturally appropriate activities (such as fishing or going out in the bush) may be probed. Reports of a "weak spirit" are met with inquiries designed to elucidate what is meant, and to quantify the extent of a respondent's "weak spirit." For example, the respondent may be asked questions like "Do you have weak spirit all day/every day?" and "What time of the day does your spirit feel the most weak?"

As a screening instrument, the HANAA aims to assist in the determination of when a person should be referred to a mental health professional for further assessment. It provides for the narrative responses to be recorded which can be helpful in-the-moment as well in-the-future when it comes to further discussion of, and "yarning" about, the specific nature of a client's presenting problem.

*This *Everyday Psychometrics* was guest-authored by Sivasankaran Balaratnasingam, Zaza Lyons, and Aleksandar Janca, all of the University of Western Australia, School of Psychiatry and Clinical Neurosciences, Perth, Australia.

Used with permission of Sivasankaran Balaratnasingam, Zaza Lyons, and Aleksandar Janca.

the basis of that information, some items from the original item pool will be eliminated and others will be rewritten. How is information about the difficulty, validity, reliability, discrimination, and bias of test items—along with information from the item-characteristic curves—integrated and used to revise the test?

There are probably as many ways of approaching test revision as there are test developers. One approach is to characterize each item according to its strengths and weaknesses. Some

items may be highly reliable but lack criterion validity, whereas other items may be purely unbiased but too easy. Some items will be found to have many weaknesses, making them prime candidates for deletion or revision. For example, very difficult items have a restricted range; all or almost all testtakers get them wrong. Such items will tend to lack reliability and validity because of their restricted range, and the same can be said of very easy items.

Test developers may find that they must balance various strengths and weaknesses across items. For example, if many otherwise good items tend to be somewhat easy, the test developer may purposefully include some more difficult items even if they have other problems. Those more difficult items may be specifically targeted for rewriting. The purpose of the test also influences the blueprint or plan for the revision. For example, if the test will be used to influence major decisions about educational placement or employment, the test developer should be scrupulously concerned with item bias. If there is a need to identify the most highly skilled individuals among those being tested, items demonstrating excellent item discrimination, leading to the best possible test discrimination, will be made a priority.

As revision proceeds, the advantage of writing a large item pool becomes more and more apparent. Poor items can be eliminated in favor of those that were shown on the test tryout to be good items. Even when working with a large item pool, the revising test developer must be aware of the domain the test should sample. For some aspects of the domain, it may be particularly difficult to write good items, and indiscriminate deletion of all poorly functioning items could cause those aspects of the domain to remain untested.

Having balanced all these concerns, the test developer comes out of the revision stage with a better test. The next step is to administer the revised test under standardized conditions to a second appropriate sample of examinees. On the basis of an item analysis of data derived from this administration of the second draft of the test, the test developer may deem the test to be in its finished form. Once the test is in finished form, the test's norms may be developed from the data, and the test will be said to have been "standardized" on this (second) sample. Recall from Chapter 4 that a standardization sample represents the group(s) of individuals with whom examinees' performance will be compared. All of the guidelines presented in that chapter for selecting an appropriate standardization sample should be followed.

JUST THINK . . .

Surprise! An international publisher is interested in publishing your American History Test. You've just been asked which population demographic characteristics you think are most important to be represented in your international standardization sample. Your response?

When the item analysis of data derived from a test administration indicates that the test is not yet in finished form, the steps of revision, tryout, and item analysis are repeated until the test is satisfactory and standardization can occur. Once the test items have been finalized, professional test development procedures dictate that conclusions about the test's validity await a cross-validation of findings. We'll discuss *cross-validation* shortly; for now, let's briefly consider some of the issues surrounding the development of a new edition of an existing test.

Test Revision in the Life Cycle of an Existing Test

Time waits for no person. We all get old, and tests get old, too. Just like people, some tests seem to age more gracefully than others. For example, as we will see when we study projective techniques in Chapter 12, the Rorschach Inkblot Test seems to have held up quite well over the years. By contrast, the stimulus materials for another projective technique, the Thematic Apperception Test (TAT), are showing their age. There comes a time in the life of most tests when the test will be revised in some way or its publication will be discontinued. When is that time?

No hard-and-fast rules exist for when to revise a test. The American Psychological Association (APA, 1996b, Standard 3.18) offered the general suggestions that an existing test

be kept in its present form as long as it remains "useful" but that it should be revised "when significant changes in the domain represented, or new conditions of test use and interpretation, make the test inappropriate for its intended use."

Practically speaking, many tests are deemed to be due for revision when any of the following conditions exist.

1. The stimulus materials look dated and current testtakers cannot relate to them.
2. The verbal content of the test, including the administration instructions and the test items, contains dated vocabulary that is not readily understood by current testtakers.
3. As popular culture changes and words take on new meanings, certain words or expressions in the test items or directions may be perceived as inappropriate or even offensive to a particular group and must therefore be changed.
4. The test norms are no longer adequate as a result of group membership changes in the population of potential testtakers.
5. The test norms are no longer adequate as a result of age-related shifts in the abilities measured over time, and so an age extension of the norms (upward, downward, or in both directions) is necessary.
6. The reliability or the validity of the test, as well as the effectiveness of individual test items, can be significantly improved by a revision.
7. The theory on which the test was originally based has been improved significantly, and these changes should be reflected in the design and content of the test.

The steps to revise an existing test parallel those to create a brand-new one. In the test conceptualization phase, the test developer must think through the objectives of the revision and how they can best be met. In the test construction phase, the proposed changes are made. Test tryout, item analysis, and test revision (in the sense of making final refinements) follow. All this sounds relatively easy and straightforward, but creating a revised edition of an existing test can be a most ambitious undertaking. For example, recalling the revision of a test called the Strong Vocational Interest Blank, Campbell (1972) reflected that the process of conceiving the revision started about ten years prior to actual revision work, and the revision work itself ran for another ten years. Butcher (2000) echoed these thoughts in an article that provided a detailed "inside view" of the process of revising a widely used personality test called the MMPI. Others have also noted the sundry considerations that must be kept in mind when conducting or contemplating the revision of an existing instrument (Adams, 2000; Cash et al., 2004; Okazaki & Sue, 2000; Prinzie et al., 2007; Reise et al., 2000; Silverstein & Nelson, 2000; Vickers-Douglas et al., 2005).

JUST THINK . . .

Why can the process of creating a revision to an established test take years to complete?

Once the successor to an established test is published, there are inevitably questions about the equivalence of the two editions. For example, does a measured full-scale IQ of 110 on the first edition of an intelligence test mean exactly the same thing as a full-scale IQ of 110 on the second edition? A number of researchers have advised caution in comparing results from an original and a revised edition of a test, despite similarities in appearance (Reitan & Wolfson, 1990; Strauss et al., 2000). Even if the content of individual items does not change, the context in which the items appear may change, thus opening up the possibility of significant differences in testtakers' interpretation of the meaning of the items. Simply developing a computerized version of a test may make a difference, at least in terms of test scores achieved by members of different populations (Ozonoff, 1995).

Formal item-analysis methods must be employed to evaluate the stability of items between revisions of the same test (Knowles & Condon, 2000). Ultimately, scores on a test and on its

updated version may not be directly comparable. As Tulsky and Ledbetter (2000) summed it up in the context of original and revised versions of tests of cognitive ability: "Any improvement or decrement in performance between the two cannot automatically be viewed as a change in examinee performance" (p. 260).

A key step in the development of all tests—brand-new or revised editions—is cross-validation. Next we discuss that important process as well as a more recent trend in test publishing, *co-validation.*

Cross-validation and co-validation The term **cross-validation** refers to the revalidation of a test on a sample of testtakers other than those on whom test performance was originally found to be a valid predictor of some criterion. We expect that items selected for the final version of the test (in part because of their high correlations with a criterion measure) will have smaller item validities when administered to a second sample of testtakers. This is so because of the operation of chance. The decrease in item validities that inevitably occurs after cross-validation of findings is referred to as **validity shrinkage.** Such shrinkage is expected and is viewed as integral to the test development process. Further, such shrinkage is infinitely preferable to a scenario wherein (spuriously) high item validities are published in a test manual as a result of inappropriately using the identical sample of testtakers for test standardization and cross-validation of findings. When such scenarios occur, test users will typically be let down by lower-than-expected test validity. The test manual accompanying commercially prepared tests should outline the test development procedures used. Reliability information, including test-retest reliability and internal consistency estimates, should be reported along with evidence of the test's validity. Articles discussing cross-validation of tests are often published in scholarly journals. For example, Bank et al. (2000) provided a detailed account of the cross-validation of an instrument used to screen for cognitive impairment in older adults.

Not to be confused with "cross-validation," **co-validation** may be defined as a test validation process conducted on two or more tests using the same sample of testtakers. When used in conjunction with the creation of norms or the revision of existing norms, this process may also be referred to as **co-norming.** A current trend among test publishers who publish more than one test designed for use with the same population is to co-validate and/or co-norm tests. Co-validation of new tests and revisions of existing tests can be beneficial in various ways to all parties in the assessment enterprise. Co-validation is beneficial to test publishers because it is economical. During the process of validating a test, many prospective testtakers must first be identified. In many instances, after being identified as a possible participant in the validation study, a person will be prescreened for suitability by means of a face-to-face or telephone interview. This costs money, which is charged to the budget for developing the test. Both money and time are saved if the same person is deemed suitable in the validation studies for multiple tests and can be scheduled to participate with a minimum of administrative preliminaries. Qualified examiners to administer the test and other personnel to assist in scoring, interpretation, and statistical analysis must also be identified, retained, and scheduled to participate in the project. The cost of retaining such professional personnel on a per-test basis is minimized when the work is done for multiple tests simultaneously.

Beyond benefits to the publisher, co-validation can hold potentially important benefits for test users and testtakers. Many tests that tend to be used together are published by the same publisher. For example, the fourth edition of the Wechsler Adult Intelligence Scale (WAIS-IV) and the fourth edition of the Wechsler Memory Scale (WMS-IV) might be used together in the clinical evaluation of an adult. And let's suppose that, after an evaluation using these two tests, differences in measured memory ability emerged as a function of the test used. Had these two tests been normed on different samples, then sampling error would be one possible reason

for the observed differences in measured memory. However, because the two tests were normed on the same population, sampling error as a causative factor has been virtually eliminated. A clinician might thus look to factors such as differences in the way that the two tests measure memory. One test, for example, might measure short-term memory using the recall of number sequences. The other test might measure the same variable using recalled comprehension of short reading passages. How each test measures the variable under study may yield important diagnostic insights.

On the other hand, consider two co-normed tests that are almost identical in how they measure the variable under study. With sampling error minimized by the co-norming process, a test user can be that much more confident that the scores on the two tests are comparable.

Quality assurance during test revision Once upon a time, a long time ago in Manhattan, one of this text's authors (Cohen) held the title of senior psychologist at Bellevue Hospital. Among other duties, senior psychologists supervised clinical psychology interns in all phases of their professional development, including the administration of psychological tests:

> One day, in the course of reviewing a test protocol handed in by an intern, something very peculiar caught my eye. On a subtest that had several tasks scored on the basis of number of seconds to completion, all of the recorded times on the protocol were in multiples of 5 (as in 10 seconds, 15 seconds, etc.). I had never seen a protocol like that. All of the completed protocols I had seen previously had recorded completion times with no identifiable pattern or multiple (like 12 seconds, 17 seconds, 9 seconds, etc.). Curious about the way that the protocol had been scored, I called in the intern to discuss it.
>
> As it turned out, the intern had not equipped herself with either a stopwatch or a watch with a second-hand before administering this test. She had ignored this mandatory bit of preparation prior to test administration. Lacking any way to record the exact number of seconds it took to complete each task, the intern said she had "estimated" the number of seconds. Estimating under such circumstances is not permitted because it violates the standardized procedure set forth in the manual. Beyond that, estimating could easily result in the testtaker either earning or failing to earn bonus points for (inaccurately) timed scores. The intern was advised of the error of her ways, and the patient was retested.

Well, that's one "up close and personal" example of quality control in psychological testing at a large municipal hospital. But what mechanisms of quality assurance are put into place by test publishers in the course of standardizing a new test or restandardizing an existing test? Let's take a brief look at some quality control mechanisms for examiners, protocol scoring, and data entry. For the purpose of illustration, we draw some examples from procedures followed by the developers of the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV; Wechsler, 2003).

The examiner is the front-line person in test development, and it is critically important that examiners adhere to standardized procedures. In developing a new test or in restandardizing or renorming an existing test, test developers seek to employ examiners who have experience testing members of the population targeted for the test. For example, the developers of the WISC-IV sought to

> recruit examiners with extensive experience testing children and adolescents. Potential examiners completed a questionnaire by supplying information about their educational and professional experience, administration experience with various intellectual measures, certification, and licensing status. Those selected as potential standardization examiners were very familiar with childhood assessment practices. (Wechsler, 2003, p. 22)

Although it might be desirable for every examiner to hold a doctoral degree, this is simply not feasible given that many thousands of tests may have to be individually administered. The professional time of doctoral-level examiners tends to be at a premium—not to mention their fees. Regardless of education or experience, all examiners will be trained to administer the

instrument. Training will typically take the form of written guidelines for test administration and may involve everything from classroom instruction to practice test administrations on site to videotaped demonstrations to be reviewed at home. Publishers may evaluate potential examiners by a quiz or other means to determine how well they have learned what they need to know. During the standardization of the WISC-IV, examiners were required to submit a review case prior to testing additional children. And during the course of the test's standardization, all persons selected as examiners received a periodic newsletter advising them of potential problems in test administration. The newsletter was designed to provide an ongoing way to maintain quality assurance in test administration.

In the course of test development, examiners may be involved to greater or lesser degrees in the final scoring of protocols. Regardless of whether it is the examiner or a "dedicated scorer," all persons who have responsibility for scoring protocols will typically undergo training. As with examiner training, the training for scorers may take many forms, from classroom instruction to videotaped demonstrations.

Quality assurance in the restandardization of the WISC-IV was in part maintained by having two qualified scorers rescore each protocol collected during the national tryout and standardization stages of test development. If there were discrepancies in scoring, the discrepancies were resolved by yet another scorer, referred to as a *resolver*. According to the manual, "The resolvers were selected based on their demonstration of exceptional scoring accuracy and previous scoring experience" (Wechsler, 2003, p. 22).

Another mechanism for ensuring consistency in scoring is the *anchor protocol*. An **anchor protocol** is a test protocol scored by a highly authoritative scorer that is designed as a model for scoring and a mechanism for resolving scoring discrepancies. A discrepancy between scoring in an anchor protocol and the scoring of another protocol is referred to as **scoring drift.** Anchor protocols were used for quality assurance in the development of the WISC-IV:

> If two independent scorers made the same scoring error on a protocol, comparison to the anchor score revealed the scoring drift. Scorers received feedback immediately to prevent repetition of the error and to correct for scoring drift. (Wechsler, 2003, p. 23)

Once protocols are scored, the data from them must be entered into a database. For quality assurance during the data entry phase of test development, test developers may employ computer programs to seek out and identify any irregularities in score reporting. For example, if a score on a particular subtest can range from a low of 1 to a high of 10, any score reported out of that range would be flagged by the computer. Additionally, a proportion of protocols can be randomly selected to make certain that the data entered from them faithfully match the data they originally contained.

The Use of IRT in Building and Revising Tests

In the previous chapter, we noted that item response theory (IRT) could be applied in the evaluation of the utility of tests and testing programs. Here, let's briefly elaborate on the possible roles of IRT in test construction, as well as some of its pros and cons vis-à-vis classical test theory (CTT). As can be seen from Table 8–4, one of the *disadvantages* of applying CTT in test development is the extent to which item statistics are dependent upon characteristics (strength of traits or ability level) of the group of people tested. Stated another way, "all CTT-based statistics are sample dependent" (De Champlain, 2010, p. 112). To elaborate, consider a hypothetical "Perceptual-Motor Ability Test" (PMAT), and the characteristics of items on that test with reference to different groups of testtakers. From a CTT perspective, a PMAT item might be judged to be very *high* in difficulty when it is administered to a sample of people known to be very low in perceptual-motor ability. From

Table 8–4
Some Advantages and Disadvantages of Classical Test Theory (CTT) and Item Response Theory (IRT)*

Theory	Advantages	Disadvantages
Classical Test Theory	1. Smaller sample sizes are required for testing, so CTT is especially useful if only a small sample of testtakers is available. 2. CTT utilizes relatively simple mathematical models. 3. Assumptions underlying CTT are "weak" allowing CTT wide applicability 4. Most researchers are familiar with this basic approach to test development. 5. Many data analysis and statistics-related software packages are built from a CTT perspective or are readily compatible with it.	1. Item statistics and overall psychometric properties of a test are dependent on the samples which have been administered the test. 2. Tests developed using CTT may be longer (or, require more items) than tests developed using IRT. 3. One often violated assumption is that each item of a test contributes equally to the total test score.
Item Response Theory	1. Item statistics are independent of the samples which have been administered the test. 2. Test items can be matched to ability levels (as in computerized adaptive testing) thus resulting in relatively short tests that are still reliable and valid. 3. IRT models facilitate advanced psychometric tools and methods, holding out the promise of greater precision in measurement under certain circumstances.	1. The techniques used to test item response models are relatively complicated and unfamiliar to most researchers. 2. Sample sizes need to be relatively large to properly test IRT models (200 or more is a good rule-of-thumb). 3. Assumptions for use of IRT are characterized as "hard" or "strong" making IRT inappropriate for use in many applications. 4. As compared to CTT-based statistics-related software, there are much fewer IRT-based packages currently available.

*For a more detailed comparison of CTT to IRT, consult the sources used to synthesize this table (De Champlain, 2010; Hambleton & Jones, 1993; Streiner, 2010; and Zickar & Broadfoot, 2009).

that same perspective, that same PMAT item might be judged to be very *low* in difficulty when administered to a group of people known to be very high in perceptual-motor ability. Because the way that an item is viewed is so dependent on the group of testtakers taking the test, the ideal situation, at least from the CTT perspective, is one in which all testtakers represent a truly random sample of how well the trait or ability being studied is represented in the population. Using IRT, test developers evaluate individual item performance with reference to item-characteristic curves (ICCs). ICCs provide information about the relationship between the performance of individual items and the presumed underlying ability (or trait) level in the testtaker.

Three of the many possible applications of IRT in building and revising tests include (1) evaluating existing tests for the purpose of mapping test revisions, (2) determining measurement equivalence across testtaker populations, and (3) developing item banks.

Evaluating the properties of existing tests and guiding test revision IRT information curves can help test developers evaluate how well an individual item (or entire test) is working to measure different levels of the underlying construct. Developers can use these information curves to weed out uninformative questions or to eliminate redundant items that provide duplicate levels of information. Information curves allow test developers to tailor an instrument to provide high information (or, precision). As an illustration, refer back to the information curve for a measure of depression in Figure 3 of the OOBAL 5-B2 (page 173). Now suppose the test developer wanted to increase precision so that level of

depression could better be measured across all levels of theta. The graph suggests that this could be accomplished by adding more items to the test (or by adding more response options to existing items) that differentiate among people with mild depressive symptoms. Adding appropriate items (or response options) will both broaden the range and increase the height of the curve across the underlying construct—thus reflecting increased precision in measurement.

Determining measurement equivalence across testtaker populations Test developers often aspire to have their tests become so popular that they will be translated into other languages and used in many places throughout the world. But how do they assure that their tests are tapping into the same construct regardless of who in the world is responding to the test items? One tool to help ensure that the same construct is being measured, no matter what language the test has been translated into, is IRT.

Despite carefully translated test items, it sometimes happens that even though the words may be linguistically equivalent, members of different populations—typically members of populations other than the population for which the test was initially developed—may interpret the items differently. As we saw in Chapter 5, for example, response rates to a measure of depression from people of different cultures may not necessarily depend on how depressed the testtaker is. Rather, response rates may vary more as a function of how much the prevailing culture sanctions outward expression of emotion. This phenomenon, wherein an item functions differently in one group of testtakers as compared to another group of testtakers known to have the same (or similar) level of the underlying trait, is referred to as **differential item functioning (DIF).** Instruments containing such items may have reduced validity for between-group comparisons because their scores may indicate a variety of attributes other than those the scale is intended to measure.

JUST THINK . . .

Create a test item that might be interpreted differently when read by younger Americans (20-something) than when read by older Americans (70-something).

In a process known as **DIF analysis,** test developers scrutinize group-by-group item response curves, looking for what are termed *DIF items.* **DIF items** are those items that respondents from different groups at the same level of the underlying trait have different probabilities of endorsing as a function of their group membership. DIF analysis has been used to evaluate measurement equivalence in item content across groups that vary by culture, gender, and age. It has even been used to explore differential item functioning as a function of different patterns of guessing on the part of members of different groups (DeMars & Wise, 2010). Yet another application of DIF analysis has to do with the evaluation of item-ordering effects, and the effects of different test administration procedures (such as paper-and-pencil test administration versus computer-administered testing).

Developing item banks Developing an item bank is not simply a matter of collecting a large number of items. Typically, each of the items assembled as part of an item bank, whether taken from an existing test (with appropriate permissions, if necessary) or written especially for the item bank, have undergone rigorous qualitative and quantitative evaluation (Reeve et al., 2007). As can be seen from Figure 8–7, many item banking efforts begin with the collection of appropriate items from existing instruments (Instruments A, B, and C). New items may also be written when existing measures are either not available or do not tap targeted aspects of the construct being measured.

All items available for use as well as new items created especially for the item bank constitute the item pool. The item pool is then evaluated by content experts, potential respondents, and survey experts using a variety of qualitative and quantitative methods. Individual items in an item pool may be evaluated by cognitive testing procedures whereby an

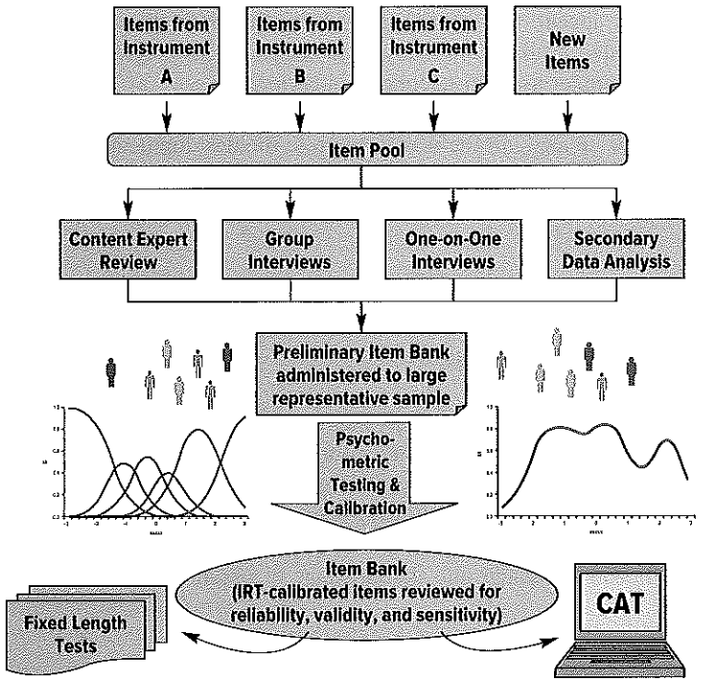

Figure 8–7
The Use of IRT to Create Item Banks

interviewer conducts one-on-one interviews with respondents in an effort to identify any ambiguities associated with the items. Item pools may also be evaluated by groups of respondents, which allows for discussion of the clarity and relevance of each item, among other item characteristics. The items that "make the cut" after such scrutiny constitute the preliminary item bank.

The next step in creating the item bank is the administration of all of the questionnaire items to a large and representative sample of the target population. For ease in data analysis, group administration by computer is preferable. However, depending upon the content and method of administration required by the items, the questionnaire (or portions of it) may be administered individually using paper-and-pencil methods.

After administration of the preliminary item bank to the entire sample of respondents, responses to the items are evaluated with regard to several variables such as validity, reliability, domain coverage, and differential item functioning. The final item bank will consist of a large set of items all measuring a single domain (or, a single trait or ability). A test developer may then use the banked items to create one or more tests with a fixed number of items. For example, a teacher may create two different versions of a math test in order to minimize efforts by testtakers to cheat. The item bank can also be used for purposes of computerized-adaptive testing.

When used within a CAT environment, a testtaker's response to an item may automatically trigger which item is presented to the testtaker next. The software has been programmed to present the item next that will be most informative with regard to the testtaker's standing on the construct being measured. This programming is actually based on near-instantaneous construction and analysis of IRT information curves. The process continues until the testing is terminated.

Because of CAT's widespread appeal, the technology is being increasingly applied to a wide array of tests. It is also becoming available on many different platforms ranging from the Internet to handheld devices to computer-assisted telephone interviewing.

Our survey of how tests are built has taken us from a test developer's first thoughts regarding what new test needs to be created, all the way through to the development of a large item bank. In reading about aspects of professional test development, it may have occurred to you that some parallel types of processes go into the development of less formal, instructor-devised measures for in-class use.

Instructor-Made Tests for In-Class Use

Professors want to give—and students want to take—tests that are reliable and valid measures of student knowledge. Even students who have not taken a course in psychological testing and assessment seem to understand psychometric issues regarding the tests administered in the classroom. As an illustration, consider each of the following pairs of statements in Table 8–5. The first statement in each pair is a criticism of a classroom test you may have heard (or said yourself); the second is that criticism translated into the language of psychometrics.

Addressing Concerns About Classroom Tests

Like their students, professors have concerns about the tests they administer. They want their examination questions to be clear, relevant, and representative of the material covered. They sometimes wonder about the length of their examinations. Their concern is to cover voluminous amounts of material while still providing enough time for students to give thoughtful consideration to their answers.

For most published psychological tests, these types of psychometric concerns would be addressed in a formal way during the test development process. In the classroom, however, rigorous psychometric evaluation of the dozen or so tests that any one instructor may administer during the course of a semester is impractical. Classroom tests are typically created for the purpose of testing just one group of students during one semester. Tests change to reflect changes in lectures and readings as courses evolve. Also, if tests are reused, they are in danger of becoming measures of who has seen or heard about the examination before taking it rather than measures of how well the students know the course material. Of

Table 8–5
Psychometric "Translation" of Student Complaints

Student Complaint	Translation
"I spent all last night studying Chapter 3, and there wasn't one item on that test from that chapter!"	"I question the examination's content validity!"
"The instructions on that essay test weren't clear, and I think it affected my grade."	"There was excessive error variance related to the test administration procedures."
"I wrote the same thing my friend did for this short-answer question—how come she got full credit and the professor took three points off my answer?"	"I have grave concerns about rater error affecting reliability."
"I didn't have enough time to finish; this test didn't measure what I know—only how fast I could write!"	"I wish the person who wrote this test had paid more attention to issues related to criterion-related validity and the comparative efficacy of speed as opposed to power tests!"

course, although formal psychometric evaluation of classroom tests may be impractical, informal methods are frequently used.

Concerns about content validity are routinely addressed, usually informally, by professors in the test development process. For example, suppose an examination containing 50 multiple-choice questions and five short essays is to cover the reading and lecture material on four broad topics. The professor might systematically include 12 or 13 multiple-choice questions and at least one short essay from each topic area. The professor might also draw a certain percentage of the questions from the readings and a certain percentage from the lectures. Such a deliberate approach to content coverage may well boost the test's content validity, although no formal evaluation of the test's content validity will be made. The professor may also make an effort to inform the students that all textbook boxes and appendices and all instructional media presented in class (such as videotapes) are fair game for evaluation.

Criterion-related validity is difficult to establish on many classroom tests because no obvious criterion reflects the level of the students' knowledge of the material. Exceptions may exist for students in a technical or applied program who take an examination for licensure or certification. Informal assessment of something akin to criterion validity may occur on an individual basis in a student–professor chat wherein a student who obtained the lowest score in a class may demonstrate to the professor an unambiguous lack of understanding of the material. It is also true that the criterion validity of the test may be called into question by the same method. A chat with the student who scored the highest might reveal that this student doesn't have a clue about the material the test was designed to tap. Such a finding would give the professor pause.

The construct validity of classroom tests is often assessed informally, as when an anomaly in test performance may call attention to issues related to construct validity. For example, consider a group of students who have a history of performing at an above-average level on exams. Now suppose that all the students in this group perform poorly on a particular exam. If all these students report not having studied for the test or just not having understood the text material, then there is an adequate explanation for their low scores. However, if the students report that they studied and understood the material as usual, then one might explain the outcome by questioning the exam's construct validity.

Aspects of a classroom test's reliability can also be informally assessed. For example, a discussion with students can shed light on the test's internal consistency. Then again, if the test was designed to be heterogeneous, then low internal consistency ratings might be desirable. On essay tests, inter-rater reliability can be explored by providing a group of volunteers with the criteria used in grading the essays and letting them grade some. Such an exercise might clarify the scoring criteria. In the rare instance when the same classroom test is given twice or in an alternate form, a discussion of the test-retest or alternate-forms reliability can be conducted.

Have you ever taken an exam in which one student quietly asks for clarification of a specific question and the professor then announces to the entire class the response to the student's question? This professor is attempting to reduce administration error (and increase reliability) by providing the same experience for all testtakers. When grading short-answer or essay questions, professors may try to reduce rater error by several techniques. For example, they may ask a colleague to help decipher a student's poor handwriting or re-grade a set of essays (without seeing the original grades). Professors also try to reduce administration error and increase reliability by eliminating items that many students misunderstand.

Tests developed for classroom use may not be perfect. Few, if any, tests for any purpose are. Still, most professors much like their professional test developer counterparts, are always on the lookout for ways—to make their tests as psychometrically sound as possible. In the following chapters, we will be exploring various aspects of many different types of tests, beginning with tests of intelligence. But before discussing tests of *intelligence*, reflect for a moment—and once again when you read Chapter 9—on the meaning of that somewhat elusive term.

Self-Assessment

Test your understanding of elements of this chapter by seeing if you can explain each of the following terms, expressions, and abbreviations:

anchor protocol
asexuality
biased test item
binary-choice item
categorical scaling
category scoring
ceiling effect
class scoring
comparative scaling
completion item
computerized adaptive testing (CAT)
co-norming
constructed-response format
co-validation
cross-validation
DIF analysis
differential item functioning (DIF)
DIF items
essay item
expert panel
floor effect
giveaway item
guessing
Guttman scale
ipsative scoring
item analysis
item bank
item branching
item-characteristic curve (ICC)
item-difficulty index
item-discrimination index
item-endorsement index
item fairness
item format
item pool
item-reliability index
item-validity index
Likert scale
LGBTA
matching item
method of paired comparisons
multiple-choice format
pilot work
pseudobulbar affect (PBA)
qualitative item analysis
qualitative methods
rating scale
scaling
scalogram analysis
scoring drift
selected-response format
sensitivity review
short-answer item
summative scale
test conceptualization
test construction
test development
test revision
test tryout
"think aloud" test administration
true–false item
validity shrinkage