CHAPTER 11 Measurement Tools and Strategies

What do I need to know to assess already prepared instruments and create strong ones of my own?

Once you decide upon what you want to measure, the strength of any scale you want to use must be evaluated. This chapter provides an overview of the concepts of reliability and validity to help you understand these important considerations.

What skills will I need to evaluate or create my own instruments?

You must be able to read the literature that describes potential instruments being considered for adoption and be able to discern which scales have the strongest evidence of validity and reliability. And when the available scales aren't really acceptable, you need to be able to begin the process of validating and establishing the credibility of a scale being developed for your unique program. You must be able to objectively evaluate instruments—including the ones you create.

WHY IS GOOD MEASUREMENT SO IMPORTANT TO EVALUATION?

There once was a primitive culture where counting followed this scheme: one, two, many. If you had more than two of something, you had many.

It is clear that our society is much more concerned with counting and measuring things. If we accept a new position paying a salary of $40,000 a year, we certainly do not want to be paid $19,000 or even $29,000. If we buy three pounds of beef at the supermarket, we do not want to be charged for five pounds. Measurement, if not accountability, is a fundamental aspect of our society.

We measure almost everything: the speed of computers, the horsepower of cars, the calories we burn jogging, the interest rate the bank or credit card company charges. We measure virtually every phenomenon because we want to know if things are changing or improving, because we are hungry for information about the world and our place in it. Progress and our ability to demonstrate and quantify it have become increasingly important to us—whether we are consumers, providers, or program evaluators.

To be accountable and show progress, precise measurements must be taken. Evaluators do not just rush out and start gathering data. As we learned earlier, outcome variables must be operationalized. On many occasions, evaluators use paper-and-pencil instruments to form the basis for measurements.

We sometimes call these instruments questionnaires, although they do not always ask questions of respondents. Some instruments are composed of a number of statements (items) to which the respondent indicates levels of agreement or disagreement. Instruments can be complex and composed of many items or as simple as three or four items. Well-developed instruments help us understand why some clients benefitted from an intervention and others did not. They also allow us to examine more closely those clients whose progress was meager or moderate—interventions do not always have the same effect on every person.

Instruments permit us to use quantification to move beyond subjective opinions ("I think these clients have improved") into a domain where we can discuss the amount of change or improvement ("This group of clients is 37 percent more assertive than they were at the time of pretest" or "At the time of the posttest, 55 percent of the intervention group reported no clinically significant symptoms").

Objective instruments provide evaluators with a certain amount of precision in arriving at the magnitude or intensity of clients' problems and in determining any consequent change in those problems. We are afforded this accuracy because rigorous procedures underlie the construction of useful instruments that then allow us to quantify abstract or intangible concepts such as self-esteem or assertiveness. These instruments allow us to translate subjective perceptions of problems and concepts into numeric values.

Well-constructed evaluation instruments are not just thrown together over an evening or weekend. However, it is easy to find scales and questionnaires on the Internet and in the literature that could provide unreliable or worthless information. How do we identify the strongest instruments and separate them from the weaker ones? That's the purpose of this chapter. It is important to select instruments that are: a) good indicators of what the programs are attempting to accomplish and b) psychometrically strong—more about this later in the chapter.

Consider the following scenario:

> Jim Gradstudent was asked to evaluate a residential treatment center for youth who had experienced trouble with the juvenile justice system. He noticed that several of the most successful residents at the time of release seemed to have experienced an increase in self-esteem. One afternoon while browsing on the Internet, he found a self-esteem instrument that looked as if it could be used for program evaluation. After deciding on a one-group pretest–posttest design, Jim made a number of photocopies of the self-esteem instrument and, with the agency director's permission, began administering it to new admissions. Eleven months later, he had collected 42 pre- and posttests from residents who had been discharged from the treatment center. He was surprised to learn

that there was very little difference in the pre- and posttest scores. Does this mean that the residential treatment program was unsuccessful?

As you think about this scenario, the lack of information in several areas should cause you to raise questions. Is increasing the resident's self-esteem a clearly articulated goal of the residential treatment center? If it is not, would use of a self-esteem inventory to evaluate the whole program be a reasonable measure? Even if it is an important goal of the program, how likely is it that the youth in treatment will actually experience an increase in self-esteem? Is there evidence from the literature to suggest that residential treatment centers for this population commonly increase self-esteem? Could any other variable be used to gauge the success of the program?

If it occurred to you that a better measure of the treatment center's success might be recidivism—subsequent arrests or offenses that would bring these youths to the attention of the juvenile justice system again—you are right. It is not always necessary to use a paper-and-pencil instrument to measure program outcome. Furthermore, the use of an instrument of unknown psychometric qualities should be avoided. Evaluators need to know how "good" instruments are. Because we know nothing about the reliability and validity of the instrument that Jim selected, it is possible that it might not have detected changes in self-esteem even if they did occur. But more about that later. First, let us examine our alternatives.

WHAT SHOULD WE MEASURE?

We start deciding what to measure when we ask how a program's "success" could best be demonstrated. What is a program trying to accomplish? If it fails, how would that failure be noted? As you can see in Table 11.1, for some programs the criteria are obvious.

Table 11.1 Examples of Behavioral Outcomes for Programs

Program	Success	Failure
Alcohol and drug treatment programs	Days of sobriety	Days of drinking, public intoxication, DUIs
Day treatment for the severely mentally ill	Days of independent living in community	Rehospitalizations, number of days in hospital
Juvenile and adult criminal justice diversion programs	Days without arrest; employment or school attendance	Rearrests; days in jail; suspension from school
Employment and training programs	Wages and hours worked	Amount of entitlements received
Adoption programs	Number of permanent placements made	Number of children eligible for adoption still in foster care
Child protection programs	No further reports or evidence of abuse or neglect	Substantiated reports of abuse or neglect

Programs can be evaluated with the data they routinely collect. It is very likely, for instance, that a mental health center will know how many or what percent of its clients in the day treatment program became hospitalized during the course of a year. To change examples, it should not be difficult for adoption workers to determine the number or percentage of children for whom a permanent placement was obtained. Forensic programs ought to be able to determine which of their clients are rearrested.

Schools know which students drop out and which ones graduate. Dropping out of school can be viewed as a behavior in much the same way as the logical consequences of substance abuse (such as arrest or DUI) reflect certain behaviors. Official records can be used to gauge the impact of interventions; programs can be evaluated in terms of *behavioral outcomes* without interviewing, observing, or distributing questionnaires to program recipients.

Many human service agencies are required to annually publish data on the number of clients they serve. Often these data are available on a county basis and may be useful to evaluators trying to determine the impact of broadly focused programs.

The use of official data to measure progress is nothing new. Florence Nightingale is said to have kept statistics on the mortality of British soldiers. She kept track of hospital deaths by diagnostic categories to show that improvements in sanitation reduced fatalities. Because of her efforts, the mortality rate dropped from 32 percent to 2 percent within six months (M. A. Nutting and L. L. Dock, 1907, cited in Meisenheimer, 1985).

Behavioral data can include such specific physiological measurements as those obtained from skinfold calipers, from measuring weight gained or lost, or from biochemical measures, such as urine analyses, to detect which clients are staying "clean." Keeping track of binging or purging episodes or number of sleepless nights are also behavioral data.

Behavioral data can be obtained through the use of client self-monitoring and calendar recall methods. For instance, Yoshihama, Gillespie, Belli, and Tolman (2005) have reported that use of a life history calendar resulted in a significantly higher report of lifetime interpersonal violence when compared to a comparison group questioned with a structured interview method.

Video- and audiotaping clients' interactions with others are additional sources of behavioral data. For instance, to determine if parents interact more appropriately with their young children after a nurturing program, videotaping sessions with parents and their children during meal times could be arranged. Although there might be some concerns with "staged" behavior, a benefit of videotaping is that facial expressions and general demeanor can be observed. Rating scales can be developed so that there is a quantitative count or rating on the presence or absence of desired nurturing behaviors. The tapes can also be used by clients for learning from self-observation, as well as for demonstrating progress.

However, behavioral outcomes are not always available to the evaluator. For instance, suppose you are the director of a program that provides drug prevention programming for elementary school children. The goal of your program is to prevent these children from becoming substance abusers as adolescents or adults. Most programs of this type will not have the ability to do any sort of follow-up study 3,

5, or 10 years later to see if the prevention programming resulted in fewer persons with drug dependency problems than in the control group. As a consequence, these and other programs without an ability to measure behavioral outcomes must consider success in terms of clients increasing their *knowledge* about a given problem or in terms of changing clients' *attitudes*.

Sometimes prevention programs measure whether the program recipients have increased their knowledge about a given problem. In HIV/AIDS prevention programs, for instance, the goal could be to provide sufficient information about how HIV/AIDS is transmitted so that program participants have an increased knowledge about its transmission. One could envision a pretest of 20 items and the typical respondent (before the intervention) getting four or five items correct. After the intervention (assuming that the educational presentation is effective), the typical respondent might answer 18 or 19 items correctly on the posttest. This would indicate that respondents' knowledge about AIDS had been increased.

For other programs, the main goal may be to change the participants' attitudes about some behavior or practice. For instance, if you were administering an intervention program for men who batter, the evaluator might use a behavioral measure (arrests, incidents of battering) as an outcome measure, but it would also be possible to determine if program participants had a change in attitudes about intimate partner violence. The goal of the program might be to help batterers become more empathetic—to put themselves in the place of the victim—and to view battering as unacceptable behavior. In this instance, the evaluator may not want to measure batterers' knowledge about domestic violence but to change attitudes regarding its acceptability. The theory here would be that if attitudes change, so will behavior.

Often, it is much easier to measure attitudes and knowledge than behavior. It is relatively easy to determine if adolescents have become more knowledgeable about drugs or if they have developed attitudes favorable to the use of illicit drugs. It is much more difficult to determine if program recipients sell, buy, or use illegal drugs once they are away from school. Using the men-who-batter example, even after a treatment program has been completed, battering may still occur in the home but go unreported. An evaluator might be tempted to conclude that an intervention program was successful because there were no rearrests among the program participants, when in reality battering was still occurring but less often or in a somewhat less severe form.

A major advantage of paper-and-pencil measures of knowledge and attitudes is that they can be administered easily in a classroom, waiting room, or office—and thus outcome data can usually be obtained more quickly than waiting for some future behavioral measure (such as clients being rearrested or hospitalized) that require weeks or months to pass.

A major disadvantage of focusing on knowledge and attitudes is that they may not be directly related to behavior. For example, clients may have knowledge that drug use is bad for them but continue using, nonetheless. (Consider individuals who smoke cigarettes even though the surgeon general's warning is printed on each pack.) Clients can increase their knowledge about alcoholism (or a number of other problems) and yet not change their behavior.

BOX 11.1 Scales, Indices, Tests

A **scale** is generally considered to be an item that measures a solitary concept (like hostility or anxiety) and commonly attempts to assess the intensity or amount of that concept. For instance, a scale might consist of a single item (i.e., "Rate how anxious you are feeling today on a scale from 1 to 10.").

An **index** involves the creation of a new variable that is the sum of other items or variables that are thought to measure a single construct. Thus, a researcher might need 25 items to get an accurate measurement of clients' anxiety. The idea is that a composite score of these items is a stronger and more robust way to measure a slippery and often intangible construct. Don't be confused, however, when you begin looking at instruments and find out that most researchers use the terms *scale* and *index* interchangeably when multiple items are used to create a single score for an individual. (Just look at the Clinical Anxiety Scale in the next chapter and you'll see what we mean.)

Tests are slang for just about any paper-and-pencil instruments that attempt to create some quantitative score.

An **instrument** may include several **subscales**—each dedicated to evaluating a different construct. For instance, see Whitt and Howard's (2012) article on the Brief Symptom Inventory (BSI). They were able to reduce the BSI's 53 items and nine symptom dimensions to 25 items and six psychiatric symptom subscales for assessing antisocial adolescents involved with the juvenile justice system.

The connection between attitudes, knowledge, and behavior is tenuous at best. Probably the "best" measure in any situation would be one closest to the intent of the program intervention. When it is not possible to observe behavioral changes or to get reliable measures of specific behaviors from other sources, evaluators often use paper-and-pencil instruments to measure changes in attitudes, knowledge, or self-reported behavior.

Research instruments can be discussed and are evaluated along two primary psychometric dimensions: reliability and validity.

RELIABILITY

An instrument or questionnaire is said to be **reliable** when it consistently and dependably measures some concept or phenomenon with accuracy. If an instrument is reliable, then administering it to similar groups yields similar results. A reliable instrument is like a reliable watch—it should not be easily affected by external factors such as temperature, humidity, day of the week, cycle of the moon, and so forth.

The reliability of instruments is generally reported in a way that resembles a correlation coefficient—it will be a numerical value between 0 and 1. Nunnally, and Bernstein (1994) say that in the early stages of research, one can work with instruments having modest reliability (by which they mean .70 or higher), that .80 can be used for basic research, and that a reliability of .90 is the minimum where important decisions are going to be made with respect to specific test scores.

What does it mean when an instrument does not have even modest reliability? It means that the instrument cannot be counted on to be consistent. In other words, its accuracy varies in ways that might not be well understood. It might provide a good measurement of some concept in one situation and be inaccurate in another situation. How could this happen? One way would be that the items are vague and not interpreted in the same way by various individuals. For instance, suppose I am interested in measuring knowledge about AIDS and I develop the following item: "It is possible to get AIDS from gay employees in restaurants or bars." Six out of ten individuals may interpret this item as asking whether food handlers can transmit AIDS, presumably by handling plates, silverware, or breathing on food. However, if four out of every ten individuals read into the item the question of whether AIDS is transmitted by having sex with the gay employees of restaurants and bars, then this item would detract from, rather than contribute to, the making of a reliable instrument.

Reliable items provide a consistent frame of reference. If an item can be interpreted in several different ways, then it should be tossed out. Pilot testing of questionnaires and data-gathering instruments on a small scale can often identify items that confuse respondents. Although a single item will seldom make a whole scale unreliable, several vague items can affect reliability. It is always in the evaluator's best interest to use as reliable a scale as is possible.

When a scale or instrument is used and reported in a professional journal article or evaluation effort, the author should include information about it. If there is no information on the instrumentation, there can be no presumption of reliability or validity. This problem commonly arises when the author's instrument or questionnaire is "homemade." Whether you or a committee design a questionnaire, it cannot be assumed to be reliable until it has been tested.

WHAT DO I NEED TO KNOW ABOUT RELIABILITY?

Although there are several ways to demonstrate reliability, most researchers start first with **internal consistency**. With this approach, each of the individual items that make up a scale is examined for how well it correlates with the scale as a whole. IBM SPSS Statistics is one of several computer software programs that can determine a scale's reliability. The reliability procedure provides an item analysis that helps the researcher know which items to drop.

To show you this process, we have incorporated data from a scale being developed to measure adolescents' attitudes about the value of work (see Box 11.2). Approximately 100 adolescents completed the "My Attitudes About Work" scale. When these data were entered into the computer, the printout in Table 11.2 was obtained.

Look at the column headed "Corrected Item—Total Correlation." Q5 and Q8 stand apart from the rest because they are negatively correlated to the scale as a whole. Including these items in the scale has the result of lowering the alpha (reliability coefficient). This outcome can be determined by looking at the column on the far right. Dropping Q5 would raise the scale's alpha to .82; deleting Q8 has

Table 11.2 Reliability Analysis—First Effort

Item	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item—Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
Q1	36.8	28.3	.55	.60	.72
Q2	37.4	25.0	.61	.50	.70
Q3	37.2	26.2	.54	.51	.71
Q4	37.1	26.2	.59	.47	.71
Q5	38.3	37.9	−.71	.75	.82
Q6	37.0	28.2	.37	.41	.73
Q7	37.1	26.4	.57	.50	.71
Q8	38.0	35.9	−.46	.29	.81
Q9	37.1	25.8	.64	.56	.70
Q10	37.2	25.5	.61	.50	.70
Q11	36.8	30.3	.24	.42	.74
Q12	36.9	28.4	.47	.42	.73
Q13	36.9	28.6	.41	.50	.73
Q14	37.1	26.3	.55	.68	.71
Q15	37.2	25.3	.64	.59	.70
Q16	37.3	27.2	.42	.36	.73
				ALPHA = .77	

BOX 11.2 My Attitudes About Work

Instructions: For each of the statements below, indicate whether it is True (T), or False (F) for you. Use a question mark (?) if you cannot decide.

_____ 1. I would like to have a full-time job someday.
_____ 2. The idea of working for a living is exciting to me.
_____ 3. If I had a job, I would expect it to be boring.
_____ 4. Working 40 hours a week in a regular job is a waste of time.
_____ 5. I would rather have a job paying minimum wage than no job at all.
_____ 6. Earning a paycheck would make me feel important.
_____ 7. Holding down a job would give me a good feeling about myself.
_____ 8. There are plenty of ways to make money without working.
_____ 9. With a job, I would have more respect for myself.
_____ 10. Any job would probably pay less than I deserve.
_____ 11. Only stupid people work for a living.
_____ 12. It is possible to enjoy one's work.
_____ 13. I want to be employed when I grow up.
_____ 14. I would rather be unemployed than have a job that pays only minimum wage.
_____ 15. I would rather be unemployed than have a boss ordering me around.
_____ 16. Working people have more pride than people who do not work.

about the same effect. Leaving them both in will produce a scale with an overall alpha of .77. Because we want the highest internal consistency possible, it would make sense to eliminate both of these items from the scale and use a shorter, revised scale.

However, let us assume that we went back over the scale, reading it closely and checking to see how the items were coded. You will note from reading Table 11.2 that the nine items (Q1, Q2, Q5, Q6, Q7, Q9, Q12, Q13, and Q16) should be coded positively because a "true" response would indicate a favorable view of work. "True" responses to items Q3, Q4, Q8, Q10, Q11, Q14, and Q15 should be coded differently (*reverse coded*) because a "true" response to these items would indicate an unfavorable view of work. In reviewing how the items were coded, instructions to the computer for items Q5 and Q8 were reverse. Once they were coded correctly, the printout in Table 11.3 resulted.

The alpha obtained the second time in Table 11.3 is higher than the software program initially estimated. This is because even though Q5 and Q8 were coded erroneously the first time, the computer had no way of knowing this and simply followed instructions—considering them as valuable elements of the scale we wanted to develop. Coded correctly, these items add to, rather than detract from, the scale, resulting in the higher reliability coefficient.

Table 11.3 Reliability Analysis—Second Effort

Item	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item—Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
Q1	38.6	54.1	.52	.60	.89
Q2	39.2	48.8	.65	.50	.88
Q3	38.9	50.4	.60	.51	.88
Q4	38.9	51.1	.58	.47	.89
Q5	38.9	49.5	.72	.75	.88
Q6	38.8	53.4	.41	.41	.89
Q7	38.8	50.8	.61	.50	.88
Q8	39.2	51.0	.47	.29	.89
Q9	38.9	50.2	.66	.56	.88
Q10	39.0	49.9	.62	.50	.88
Q11	38.6	56.4	.25	.42	.89
Q12	38.7	53.9	.48	.42	.89
Q13	38.7	54.0	.44	.50	.89
Q14	38.8	50.0	.66	.68	.88
Q15	39.0	49.3	.68	.59	.88
Q16	39.0	52.0	.45	.36	.89
				ALPHA = .89	

A developer of a scale will probably compute its internal consistency on multiple occasions, revising items and trying to find the best combinations of items and the highest alpha that can be obtained with the fewest items.

There are several other ways to determine whether an instrument has internal consistency. The **split-half technique** involves dividing a scale in half (using either top and bottom or even and odd items) and examining how well the two halves correlate with each other. Another approach is to devise **parallel** or alternate versions of the scale and administer the forms to similar groups. Reliability would be demonstrated when both versions correlate with each other—the higher the correlation coefficient, the stronger the reliability. Both of these approaches are less popular than simply computing the internal consistency with the aid of statistical software.

Still another form of reliability (**test–retest**) is demonstrated when the scale holds up well when administered to the same group of individuals on repeated occasions. Without the benefit of intervention, groups of individuals with any given problem (e.g., low self-esteem) should not experience major increases or decreases. Should an instrument show fluctuations in scores over a period of 2 to 6 weeks, for example, then one possible explanation may be that the instrument does not have strong reliability. This is an important form of reliability to consider when the evaluator is using posttests at some distance in time from the pretests or baseline measurements.

Instruments with no or very low reliability are, for all practical purposes, worthless. This is not to say that you cannot obtain extremely valuable information from a questionnaire with nothing known about its reliability. The problem is that without evidence that the items are reliable, there is no way to guarantee that the same results would be produced if they were administered again.

Survey questionnaires pose a special problem because the reliability of single items cannot be computed. However, what you can do is to examine the literature for items that have been used and refined over the years by such survey-conducting organizations as the Roper, Harris, and Gallup polls.

As a rule, adding items to a scale increases its reliability. All else being equal, a longer scale stands a better chance of having acceptable reliability than a scale of two or three items.

Novice evaluators are often tempted to develop scales of their own. However, Thyer (1992) has made this recommendation:

> Avoid this temptation like the plague! The design and validation of a scale or survey is a major project in and of itself, and a program evaluation is NO PLACE to try to develop such a measure. If you ignore this advice and prepare your own scale, your entire study's results may be called into question because the reader will have no evidence that your new measure is a reliable or valid one. As a social work journal editor, I can attest to the fact that this is a major reason why manuscripts get rejected: a well-meaning practitioner constructs his or her own idiosyncratic scale, obtains interesting results, and tries to publish them. Invariably the reviewers note this fact, reject the study, and sadly suggest that next time the writer should use a previously published scale or outcome measure with well-established reliability and validity. (pp. 139–140)

Without doubt, instances will arise when you will have to develop your own scales. There may not be an instrument available to measure the dimension you need to quantify. Or, the only instrument you find may have some problem associated with it, such as requiring a reading comprehension level that is too high, containing too many items

for children or adolescents with short attention spans, or having weak or unknown reliability and validity. When you must create your own instrument, remember that it is your burden to demonstrate that it is reliable and valid.

The Reliability of Procedures

Reliability is a concern not only if you revise or devise an instrument to use in your evaluation, but also if you rely on secondary data like rearrests, suicides, and subsequent reports of abuse and neglect. The concern this time is not with the items used to create scales or questionnaires but with the reliability of the data-gathering and reporting procedures.

Though a bit dated, a good example of how problems with recorded data can lead to erroneous conclusions comes from Siefert, Schwartz, and Ortega (1994) who investigated infant mortality rates in Michigan's child welfare system and found a higher postneonatal death rate among those infants in foster care placement than was occurring statewide. However, when the authors sought to verify information held by the Michigan Department of Social Services, they found that 5 of the 66 infants initially identified as having died during the study period were actually alive. The management information system used by the department was previously found to have a 30-percent error rate in recording entry and exit dates from specific types of placements (Lerman, 1990, cited in Siefert et al. 1994). Obviously, it is important to have accurate data that flow from standardized procedures (e.g., everyone using the same definitions and forms, and reporting on a timely basis in the same way). Individuals who are not properly trained or supervised sufficiently may not provide data in a consistent and uniform manner. Incomplete and carelessly filled out forms almost always raise concerns about the reliability of the data being examined.

Most, if not all, social indicators vastly underestimate the true incidence of social problems in our country. We know, for example, that the incidence of domestic violence, child abuse, date rape, and so forth is much greater than the records or arrests might show. Sometimes clients cross state lines and commit offenses that authorities and evaluators may not know about. And there can be indications of severe problems (e.g., suicide attempts) that never come to the attention of record-keepers or authorities.

This is not to suggest that you should refrain from using official reports to judge the success of the programs you are evaluating. In some agencies, hospitals and health care settings in particular, records may be excellent and very useful to the evaluator.

If your measurement strategy is to use existing records for the evaluation, you need to be familiar with the procedures that generated the data. You may find problems of over- or underreporting, that the staff on one shift are more conscientious or more lax than those on another, or that there were policy or procedural changes in reporting during the study period that no one remembered to tell you. Understand that any official records are somewhat limited in their ability to describe what is "really" going on, but they often constitute the best information available. In some instances, this type of data is superior to asking clients or their partners about subsequent acts of violence or participation in illegal activities. The literature in your specific field may indicate whether self-reported data from clients will provide more reliable estimates of the behavior than official records. See Box 11.3.

BOX 11.3 How Reliable Are Self-reports of Drug Use, Sexual Behavior, and Use of Treatment Services?

Questions sometimes arise about the reliability of information provided by clients in a test–retest situation—especially those with substance abuse problems, dually diagnosed, and so on. However, a number of studies suggest that, in the main, self-reports from clients are generally reliable. Here's a small sampling of studies:

- Research subjects were consistent in answering whether they had ever used alcohol, at what age they began drinking, and on how many days in the last 30 they had consumed alcohol. The study supports the reliability of self-report measures concerning alcohol use (Johnson, Pratt, Neal & Fisher, 2010).
- Self-report documentation among dually diagnosed samples is quite reliable. (Houck, Forcehimes, Gutierrez, & Bogenschutz, 2013)
- An online survey of young adults on the quantity of cigarettes they smoked is comparable to a national household survey interview (Ramo, Hall, & Prochaska, 2011).
- There was 80 percent agreement on the number of self-reported prison terms gathered through the Life Event Calendar (Sutton, Bellair, Kowalski, Light, & Hutcherson, 2011).

If the data you are collecting for your program evaluation comes from judgments, observations, or interviews of two or more persons who are rating the behavior of participants independently, then you must be concerned with **inter-rater reliability**. Suppose you are screening persons with chronic mental illness for entry into a new program that will provide them with supervised employment and an apartment. The program is much in demand and there is a long waiting list. When you meet with Jill, you immediately notice that she does not make eye contact with you and seems to be staring into space a good deal of the time. Jill appears to be distracted, preoccupied with her own inner world, and you decide she would not be a good candidate for the program.

However, your colleague, Dr. Perceptive, is not troubled by Jill's lack of eye contact, viewing this as only shyness or fear of rejection. Based on Jill's responses, Dr. Perceptive gives Jill a high rating and recommends that she be selected for the program. If you and Dr. Perceptive do not agree at least 70 percent of the time or your scores do not correlate at least at .70 on a large sample, then you do not have adequate inter-rater reliability. The major way to improve inter-rater reliability is through training and role-playing so that the raters begin to adopt a more uniform perspective and recognize the same criteria.

WHAT DO I NEED TO KNOW ABOUT VALIDITY?

The second important dimension when evaluating instruments is validity. An instrument is said to be **valid** when it closely corresponds to the concept it was designed to measure. Let's say that you are developing a self-esteem inventory and in a sudden flash of inspiration it occurs to you that a high level of self-esteem would be

indicated if respondents could identify the 27th president of the United States. If you incorporate a number of similar items into your self-esteem inventory, you probably would not create a self-esteem scale but rather a scale that measures knowledge of American history. This scale would not be valid for measuring self-esteem.

There are various ways to go about demonstrating that an instrument has validity. Sometimes experts are asked to review it to see if the entire range of the concept is represented in the sample of items selected for the scale. This is known as **content validity**. For instance, if you were developing a scale to measure progress in the treatment of bulimia and did not include the behaviors of eating uncontrollably, binge eating, or intentionally vomiting, then you would not have covered the entire range of behaviors that ought to go into a scale designed to measure progress in treating bulimia. The term **face validity** is used when one's colleagues (or other knowledgeable persons) look over an instrument and agree that it appears to measure the concept. Neither content nor face validity is sufficient for establishing that the scale has "true" validity.

The developer of a new instrument must be concerned with more than face or content validity and must amass evidence that the scale really does measure what he or she intended. **Criterion validity** means that the instrument can be validated by an external criterion. If, for example, a new scale being developed is to measure social support, then one appropriate criterion might be the number of one's close friends. Logically, someone who scores on the high range of social support should self-report more friends than individuals scoring in the lower ranges. If the scale were being prepared for use with middle school or high school students, the criterion might be teachers' estimates of the number of friends that a sample of their students had. The criterion could also come from parents who could be asked to contribute data on the number of their children's friends.

This example shows that the creator of an instrument is not "locked" into using a specific criterion to measure the validity of an instrument. However, sometimes there may be only one appropriate criterion, for instance, academic success probably is better reflected by grades/grade point average or graduation rates than any other measure.

The best external criterion may not always be easy to select, as might be the case when attempting to select one to help validate a self-esteem scale. Sometimes no single behavior best characterizes the concept. In such situations, a researcher might use a scale from the literature or prior research that has been shown to be valid for the external criterion.

Criterion validity is generally categorized as either **predictive** (of future behavior or performance) or **concurrent** (which means the ability to predict current status). Concurrent validity involves administering the new scale to the subjects, along with another scale that previous studies have shown to be a valid measure of the same concept. If scores from the two scales correlate well, then the new scale is said to have concurrent validity.

Another form of validity is known as **construct validity**. Construct validity is concerned with the theoretical relationship of the scale to other variables. It involves the testing of presumed relationships and hypotheses. Sometimes this involves the **known-groups technique** where the investigator administers the

instrument to two very different kinds of groups expecting to find major differences in the way they respond. For instance, suppose you develop an instrument to measure attitudes about drug usage. You might administer it to persons recently arrested for possession of drugs and in jail, or those beginning an outpatient drug treatment program. A contrast group could consist of persons of similar age and background who do not use drugs. If there are statistically significant differences in the means of these two groups along expected lines (e.g., the persons who admit to frequent drug use endorse a greater number of pro-drug attitudes than those who don't have antidrug attitudes), then evidence of construct validity is shown. Instruments that cannot make discriminations between two markedly different samples would be of no use to program evaluators.

For that reason, we were very interested in whether a sample of adolescents judged to have "good" attitudes about work would have higher scores than adolescents judged to have "poor" attitudes, when tested on the "My Attitudes About Work" scale discussed earlier in the chapter. Fortunately, we found statistically significant differences. Those teens who were rated as having "good" attitudes about work had higher scores on the scale than teens who were rated, by persons who knew them, as having "poor" attitudes about work. Thus, the pilot study found some beginning evidence for the validity of the instrument.

Like reliability, there are many forms and approaches to establishing validity. Factor analysis may be used to understand or confirm the structure of a scaled construct. For example, Chao and Green (2011) conducted factor analyses to examine the factors (potential subscales) within their data as they worked to devise a multiculturally sensitive mental health scale. In factor analysis, statistical procedures produce factor loadings not unlike correlations that identify items that relate and cluster around major concepts contained within the instrument. They reported, for instance, that the item "I tend to be nervous" loaded .67 on the Anxiety factor but –.21 on the Depression factor. In other words, that item correlated better with the cluster of the other four items measuring anxiety than it did with the items associated with depression. This may seem obvious, but research participants can interpret items differently than what the original researchers had planned—resulting in items sometimes being discarded because they don't load on the factors they were assumed to support.

One problem with understanding validity as a concept is that there is no standardized taxonomy of terms. As Koeske (1994) has observed:

> After nearly four decades of methodological scholarship on measurement validation and its applications in research contexts, there exists no fully comprehensive language system for identifying and differentiating types of validity and procedures for their assessment. (p. 45)

Think of the problem of establishing validity as a gradual, ongoing, confirmatory process that builds on each study. However, if you are reviewing scales for potential use in a program evaluation, you can afford to be a little more critical. Choose scales that have the most extensive evidence of validity—whether criterion, construct, convergent, predictive, or concurrent.

Although reliability and validity have been presented as separate concepts, they are interrelated in a complex fashion. If an instrument can be empirically demonstrated to have validity (and we are not talking about just face or content validity), then it can

BOX 11.4 Considerations for Selecting Outcome Measures

Instruments should be:

Relevant and appropriate to the client group. Instruments may be inappropriate in terms of clients' primary symptoms or problems, attention span, reading level, and may not correspond well with the purpose of intervention.

Easy to administer. Overly complex instruments or those with complicated instructions may not get administered uniformly across various sites—particularly if you are dependent on others to collect your data. Will it be a burden on clients and interfere with treatment?

Useful and easy to interpret. The scores you obtain should be clear and understandable—unambiguous for the majority of clients. Will the scores assist with diagnosis or treatment planning? Interpretation is often facilitated if there is a single outcome score and if norms are available on similar treatment groups and nonclients.

Reliable and valid. Additionally, is it easy to "fake" desirable scores?

Sensitive to change. If clients get better or worse, is the instrument capable of showing small gradations of improvement or deterioration? Scales with too few items may have a difficult time showing change in small increments. At the same time, you do not want a scale so sensitive that you are led to erroneous conclusions by a client having a single "bad" day.

Relatively inexpensive. Cost considerations include purchase of the instrument, scoring it, and training staff in its use.

generally be assumed to have adequate reliability. However, a reliable instrument may not be valid for the purpose we want to use it. That is, an instrument designed to assess depression in children may not provide valid measurements of depression in adults. An instrument created for respondents in one culture may not translate well in another language, which would raise concerns about the validity of any findings.

Both reliability and validity ought to be demonstrated as evidence that an instrument is psychometrically strong. This is not an either-or choice. The evaluator should try to obtain information about the instrument's reliability and validity before adopting it. If you know nothing about the reliability and validity of an instrument, it is important to realize that the results obtained from its use will have very little meaning. One obvious way of avoiding having to establish that your new scale has reliability and validity is to use instruments that already have been demonstrated to have reliability and validity in other studies.

Although our focus has been chiefly on reliability and validity, there are a number of other characteristics of instruments to consider when choosing the best scale for a program evaluation. These considerations have been summarized in Box 11.4.

HOW DO I FIND AN INSTRUMENT FOR MY EVALUATION PROJECT?

A very helpful book you may wish to obtain for your own library is *Measures for Clinical Practice and Research* by Joel Fischer and Kevin Corcoran (2013). This book, in two volumes, contains the actual scales and descriptive information for

close to 500 rapid assessment instruments. It is a fine place to start an instrument search.

Another, but older reference book, The *Handbook of Psychiatric Measures* by Rush, First, and Blacker (2007), provides information (e.g., goals, description, administration time, reliability, validity, clinical utility, and purchase information) on a variety of diagnostic measures, including quality of life, mental health status, client satisfaction, stress and life events, family risk factors, suicide risk, and child and adolescent screening measures. One problem with the *Handbook*, however, is that the reader does not always get to view the actual instruments. A CD-ROM is provided and contains about 150 of approximately 270 instruments.

Outcomes Assessment in Clinical Practice by Sederer and Dickey (1996) contains information on 18 instruments, including the Short-Form Health Survey (SF-36), the Behavior and Symptom Identification Scale (Basis 32), the Addiction Severity Index, the Global Assessment of Functioning Scale (GAF), the Life Skills Profiles, the Brief Symptom Inventory, the Eating Disorder Inventory, the Child Behavior Checklist, the Beck Depression Inventory, the Brief Psychiatric Rating Scale, the Family Burden Interview Schedule, the Quality of Life Interview, and the CSQ-8, among others. Actual specimen copies are provided for a number of these measures.

A wide variety of useful scales (usually about 25 items in length) can be purchased from Walmyr Publishing Company (**www.walmyr.com**). Many of these scales were developed by the late Walt Hudson, a prominent social work researcher, and can be purchased in blocks of 50 for approximately $25. A few examples of the scales available from Walmyr:

Clinical Anxiety Scale
Global Screening Inventory
Generalized Contentment Scale
Index of Self-Esteem
Index of Peer Relations
Index of Alcohol Involvement
Index of Marital Satisfaction
Index of Sexual Satisfaction
Index of Family Relations
Index of Parental Attitudes
Index of Sexual Harassment
Index of Homophobia
Child's Attitude Toward Mother
Child's Attitude Toward Father
Index of Brother Relations
Index of Sister Relations
Children's Behavior Rating Scale
Client Satisfaction Inventory
Partner Abuse: Nonphysical
Partner Abuse: Physical
Brief Adult Assessment Scale
Index of Job Satisfaction

Test Reviews Online, can be found at **http://buros.unl.edu/buros/jsp/search.jsp**. Test Reviews Online claims to have a database of 4,000 commercially available tests and about half of them have been reviewed by the Buros Institute of Mental Measurements. Reviews of specific tests can be purchased for $15.

If you do not find the instrument you need after examining some of the resources listed above, do not despair. The next step is to conduct a thorough search of the literature. This is generally a good idea anyway, even if you have an instrument, because you need to know what others have learned when they evaluated programs similar to the one you wish to evaluate. They may have discovered that the instrument did not provide the kind of information they had hoped for, or

revisions to the instrument could have been reported and perhaps a shorter or more reliable version has been developed.

Typically, it is useful to start a search with an abstracting database such as *PsycINFO*, *MedLine*, or *EBSCOhost*. Be prepared for thousands of citations. You may want to consult with a reference librarian before you begin your literature searches. One strategy is to use one search box for the concept you are interested in (e.g., PTSD) and use a second box to indicate that you want a "scale." Searching for these two terms in the title will often result in more accurate "hits" than searching under the subject. If "scale" doesn't produce anything useful, then try PTSD and "measurement" or PTSD and "instrument" or "assessment."

Because of space limitations, few instruments are printed in journals any more. More typically, what is found are examples of items from the scale and information about the scale's reliability and validity. However, sometimes when an instrument was introduced a number of years ago, you may find the complete instrument in an early publication found in the article's reference list. It can also be profitable to find recent articles about the instrument by entering the original author or the title of the first journal article in *Web of Science*. *Web of Science* will then indicate all of the articles that have contained the original study in their reference lists. Becoming familiar with the literature helps the evaluator ground the evaluation effort in terms of theoretical models and expectations for the program's potential success rate. This is particularly important in those instances where programs have been rapidly implemented with little prior planning or design.

Another strategy is to contact faculty members who are active researchers and have expertise in the area in which you are trying to locate instruments. They may have files of scales or be able to refer you to sources or other researchers who will be able to help.

When every effort has been made to locate appropriate instruments and none have been found, then it may be time to consider developing your own instrument. The next section will provide some explanation for designing questionnaires and interview schedules.

HOW DOES ONE CONSTRUCT A "GOOD" EVALUATION INSTRUMENT? WHAT DO I NEED TO KNOW ABOUT QUESTIONNAIRE DESIGN?

First step—The evaluator needs to consider what exactly is needed to evaluate the program. The choices generally involve the dimensions of:

- knowledge (e.g., "What are three ways of constructively channeling anger?")
- behavior and symptoms (e.g., "How often do you binge?")
- attitudes, beliefs, opinions (e.g., "Children should be seen and not heard.")

The easiest data collection instruments to create are those that have a single focus. For instance, it is a lot simpler to design a client satisfaction instrument for a child care program for single mothers than to determine what needs to change in a community action agency with very low rates of program participation. Still, it is not uncommon for evaluation instruments to tap several dimensions.

To take an example, in a community health program designed to inform low-income women about the risks of smoking, you might want to measure both their knowledge and attitudes about smoking, but it will be even more important to examine the actual number of cigarettes smoked before and after the educational intervention. You might also want to see if they have behavioral intentions to enter a smoking cessation program, if they would want a referral to such a program, or even if they have shared educational information from the program with other loved ones or family members. Other key questions might come to mind, too, such as "How many times have you tried to stop smoking?" and "What method(s) have you used to try and stop smoking before?"

Second step—While it may be obvious to some, it is worth mentioning that as an evaluator you have a choice of administering your data collection instruments in several modes. Perhaps most commonly we design questionnaire scales that are *self-administered* or *self-reports*. These can be handed to clients as they wait for their appointments or at the end of a treatment session. Questionnaires that can be mailed out to potential respondents are also examples of self-reports. However, at times, characteristics of our target population (e.g., children) lead us to develop questionnaires where the respondent is *interviewed or observed*. Because personal interviews are a very expensive way to collect data, evaluators also have the option of the somewhat less expensive approach of *telephone interviews*. Increasingly, *electronic surveys* and *questionnaires* emailed to potential respondents are also options for evaluators.

Third step—Once a determination is made about the content of the scale or questionnaire, then the evaluator will need to create a pool of items for possible use. Several colleagues or a small committee could also be tapped for assistance with this step. However, be forewarned that there is truth in the old saying that "too many chefs spoil the broth." That is, the evaluation project can be pulled in different directions by the inclusion of contributors who have their own unique ideas about the project. Strive for consensus. Don't try to do too much. It is better to have a small, clean data set that is minimally intrusive in the lives of clients than one that is overly ambitious, burdensome, and lacking in cohesion and focus. Generally speaking, the longer the data collection instrument, the more the respondent may become fatigued, and the lower the response rate. A 20-item questionnaire will be better received by clients than a 200-item one. But length should be servant to the purpose and psychometrics of the instrument.

Along with designing the items, you will need to give thought to the response set. For instance, you might decide to go with a simple precoded (**closed-ended**) response set format such as:

- **Yes**
- **No**
- **Don't know**

Or, you may wish to use the balanced five-point standard **Likert** response set:

- **Strongly Agree**
- **Agree**
- **Undecided**

- Disagree
- Strongly Disagree

Providing multiple response choices (whether you use 5- or 7-point scales) help to reduce measurement error and bias. They also make it quicker and easier to analyze the data. Another benefit is that they usually are preferred by respondents as closed-ended items appear to require less time (or thinking) than **open-ended questions**.

Open-ended questions allow respondents to present their own views and explanations. Asked, "How long did you have to wait for your first appointment?" A respondent might write, "I had to call six different times. Five times I was told to call back later. The last time I was put on hold for about 10 minutes. When the receptionist finally got back to me, she was laughing and someone in the background was talking loudly about the annual office party." Just a few responses like this are very informative. Even if you decide to use closed-ended items, you ought to include one or two open-ended questions to allow respondents to tell you things that you might not know or suspect. (For an example, see Figure 7.3 in Chapter 7, on client satisfaction.)

Fourth step—Refining the data collection instrument is the next step in the design process. There are many issues to be considered before concluding that the instrument is complete. For instance, you might want to check such issues as:

- Question sequencing—ask general questions before specific ones; keep items on the same topic together. Start the questionnaire with easy items; place the more difficult items at the end.
- Difficulty level—avoid technical language, jargon, and abbreviations that are not well known. Do not use vocabulary items that only college graduates would know. Keep sentences short. Don't crowd the questionnaire with too much text.
- Personal/private information—many people dislike providing their age, income level, ethnic group, education, and similar information. Typically, it is best to ask for this information at the end of the instrument. Individuals are more likely to respond to sensitive items once they have become involved in the process of completing the questionnaire (or in the case of interviews, established rapport with the interviewer).
- Memory—don't ask questions where the respondents would not be expected to have an accurate recall.
- Length and appearance—too many items will fatigue the respondent. Don't make the instrument look like it will require a lot of work or effort to complete. It should be clean and professional in appearance.
- Cultural sensitivity—When translating from English into another language, another translator is needed to back-translate the document into English again to ensure accuracy. Even then, complex ideas may create flawed items. Willis and Zahnd (2007) found that 100 percent of monolingual subjects in their study had trouble with this question: "Was there ever a time when you would have gotten better medical care if you had belonged to a different race or medical group?" Some respondents had not received medical care in the United States, while some respondents had received medical care, but from an Asian

health service or Korean doctor where there was no issue with discrimination. Further, they found that the Korean translation did not appear to distinguish adequately between "good" and "fair."

WHAT ARE SOME OF THE COMMONS ERRORS MADE IN DEVELOPING QUESTIONNAIRES?

Developing good instruments requires much more explanation than an introductory evaluation textbook can provide. However, it is possible to point out some errors often made when constructing instruments or developing questionnaires. Please note, however, that the following illustrations do not constitute an exhaustive listing of all the ways it is possible to write bad items for questionnaires. There must be thousands of ways to do that—including the use of the wrong or misspelled words ("piers" for "peers"), sloppy proofreading that does not catch omitted words ("Listed below are a number activities" instead of "Listed below are a number *of* activities"), as well as incorrect grammar. Proofread and revise, proofread and revise. Make sure that cover letters, instructions to research participants—any piece of written communication for which you are responsible—are as clearly written as possible.

Look at the examples of questions in Box 11.5. Identify what is wrong with these questions before reading the explanation.

In the first question, notice that there are two positive evaluation choices ("excellent" and "good"), but only one negative possibility. Respondents have two opportunities to say something good about the program but only one to indicate dissatisfaction. This response scale is not balanced but biased toward positive feedback about the program. A better way to handle this would be to provide the response categories of "excellent," "good," "undecided," "fair," and "poor."

The problem with the second question is that "often" is not defined. What does often mean to you? Once a week? Once a month? Daily? The same difficulty would exist if the term *regular* were used (e.g., "Do you attend AA meetings regularly?").

The third question does not provide a mutually exclusive response. One could be single because one had never married, because one was a widow or widower, or because one had been married and was in the process of legally dissolving it. On some occasions, it may be important to list as a separate response those who are "separated."

In the fourth question, there is a problem with the response set. Note that the response categories also are not mutually exclusive. If one had been a client for exactly 6 months, both "a. 6 months or less" and "b. under a year" would be correct. There is also a problem with overlapping response categories in item 5. A client with a $20,000 income might select response "a." because it was the first category he or she read, or because it suggests the status of a higher income category. An additional problem with the income question is that "income" is a vague term. Is the intent of the question to identify the principal wage earner's annual salary? Or does the question seek to know the total family income from all sources? Also, is the question asking for take-home (net) or gross pay?

BOX 11.5 Examples of Poorly Constructed Questions

1. Please rate the quality of our services:
   a. excellent b. good c. poor
2. Do you come here often for help?
   a. yes b. no c. don't know
3. What is your marital status?
   a. single b. married c. divorced
4. How long have you been a client with us?
   a. 6 months or less b. under a year c. 1 year or longer
5. What is your income?
   a. $10,000 to $20,000 b. $20,000 to $30,000 c. $30,000 or more
6. Do you not make a practice of shopping only on weekends?
   a. yes b. no c. undecided
7. Do you have a male relative and a female relative over 55 years of age living at home with you?
   a. yes b. no c. undecided
8. Approximately how many minutes do you dream each evening?
   a. under 15 b. 16 to 30 c. more than 31 minutes
9. Wouldn't you agree that clients should keep their accounts current with the agency?
   a. yes b. no c. undecided
10. Are you an alcoholic?
    a. yes b. no c. undecided
11. People make fun of others through text messages and pictures.
    Strongly Agree Agree Undecided Disagree Strongly Disagree
12. I have been made fun of or called names by someone in a text message.
    Strongly Agree Agree Undecided Disagree Strongly Disagree

Question 6 creates problems because the word "not" makes the question more complex than it needs to be. Many people will have to read the question a second time. Some individuals will inadvertently fail to see "not." Also, note that "shopping" is not defined. Does shopping refer to all shopping—shopping for essentials as well as nonessentials? What if one runs out of milk and stops to pick up a quart on the way home from work Friday evening? Is stopping to buy a newspaper or a magazine considered to be shopping?

Item 7 is called a **double-barreled** question. It asks two things in one sentence. It is entirely possible to have a male relative over the age of 55 living at home without having a female relative over 55 residing there—and vice versa. How would you respond to this question if you had only the male relative 55 or older but not the female living at home?

Item 8 asks for information that the respondent cannot be reasonably expected to have. Most of us do not know how long we dream each evening. This question asks for information that can only produce guessing. Absurd questions asking for information that respondents do not have yield worthless data, and may result in respondents refusing to continue any further with the interview or the questionnaire.

Item 9 is an example of a leading question. Few people tend to disagree with a question that suggests the answer. Further, there is an issue here of **social desirability**. Most people do not disagree with normal social conventions (e.g., cleanliness, being sober on the job). We all want to be liked by other people, and we tend to give responses that are "acceptable" even if that is not what we really believe or how we really act.

In regard to item 10, it may not be easy for clients whose behavior is excessive or outside of "normal" social behavior to admit the true extent of their problem. For example, few active alcoholics will admit to being an alcoholic—yet they might admit to "occasionally drinking more than they should." Terms such as *alcoholic*, *junkie*, *addict*, and *delinquent* are stigmatizing to respondents, and most individuals will not deliberately choose a response that characterizes them as being flawed, deviant, or markedly different from the rest of humanity.

Additionally, the tenth question assumes that the respondent has an accepted definition of *alcoholic*—knowledge that he or she may not have. When asking questions that have the potential for forcing negative labels on respondents, it is almost always better to rephrase and ask more neutrally about the behavior itself. In this case, improved questions might be:

On how many days of the past 30 days have you had an alcoholic drink?
On how many occasions have you tried to quit drinking but were unable to do so?
Once you start drinking, do you find it difficult to stop before becoming completely intoxicated?
In the past 30 days, how many times have you taken a drink the first thing in the morning?

The eleventh item is not terrible although it could be considered double-barreled. It is shown here because the author of the item (a student) was asked to develop a set of items that would measure self-esteem. With this item I believe she was really trying to measure knowledge of cyber bullying, possibly with the hypothesis that it could affect one's self-esteem. With this item she lost her focus and confused her purpose in writing items.

Similarly in item 12, the student is measuring whether the respondent has been cyber bullied with text messages. While cyberbullying surely affects self-esteem, the item does not measure one's self-esteem. These last two items show the importance of single-mindedness when writing items. Don't get ahead of yourself and try to do too much. Let each item measure a single construct.

Once you (and/or the evaluation committee) are happy with a draft of the data collection instrument, then the next step is to conduct a **pilot study** with a small sample of the target population. This pretesting of the instrument allows you to identify any problems these respondents have with instructions or completion of the data collection instrument.

Figure 11.1 shows how difficult preparing an instrument can be. Consider the information that these five questions will produce. Will they provide the kind of evidence that will convince the hospital administrator of the need for an emergency room social worker? How can these questions be interpreted or misconstrued?

Example 1 But It Is So Easy to Design a Questionnaire

Susie Caseworker was employed as a hospital social worker in a rural community. She was one of two social workers responsible for patients in the hospital. Although Susie liked her job, one annoying problem was that the emergency room staff could page her and she would have to drop what she was doing and race to the emergency room. She was constantly being interrupted and taken away from her patients to be of assistance in the emergency room. In her opinion, this happened frequently enough to justify the hospital hiring another social worker solely for assignment in the emergency room. She discussed this with the hospital administrator, who said that he would make a decision once she had documented the need for an emergency room social worker. The five questions in Figure 11.1 are those that Susie prepared as part of that effort. Her intention was to give the survey to each nurse and physician who worked in the emergency room.

Figure 11.1 Emergency Room Survey

Place an "x" by the answer that best corresponds to your thinking.

1. There are times when a social worker could be utilized in the emergency room.
   ( ) never ( ) seldom ( ) occasionally ( ) frequently ( ) always
2. When I worked with a social worker, he/she acted in a professional manner.
   ( ) never ( ) seldom ( ) occasionally ( ) frequently ( ) always
3. When I needed a social worker, one was readily available.
   ( ) never ( ) seldom ( ) occasionally ( ) frequently ( ) always
4. I see cases where family members are not coping well with a relative's illness or injury.
   ( ) never ( ) seldom ( ) occasionally ( ) frequently ( ) always
5. I have seen situations in the emergency room where social workers could have done counseling.
   ( ) never ( ) seldom ( ) occasionally ( ) frequently ( ) always

A potential problem with the first question is that it assumes that physicians and nurses know how often and on which occasions a social worker is needed. If physicians and nurses do not know exactly what it is that a social worker does, then it is entirely possible that they would under- or overestimate the number of occasions when a social worker could be appropriately employed. Do they think social workers are to be used to hold the hand of a person in pain? Are they to provide grief counseling only when the chaplain is not around? Are social workers to watch small children when there is no one else to supervise them? Better information might be obtained if the emergency room staff were asked to identify the needed activities to be performed by social workers or the occasions when a social worker could be used.

A related but missing question could be developed to identify times that a social worker was most needed. There may be shifts (such as between 11:00 P.M. and 7:00 A.M. on weekends) when there are more emergencies requiring assistance from a social worker. It may be that the existing hospital social workers can adequately cover the emergency room during weekdays, but that the greatest need for a social worker is on weekends and evenings.

It might also be helpful to ask the emergency room staff to enumerate the number of times during an average day, weekend, and evening shift when the services of a social worker would be beneficial. Here, too, the response set is important. Knowing that respondents indicated that a social worker could have been used an average of 25 times per shift is a lot more powerful information than knowing the most frequent response was "occasionally" or "frequently."

The second question inappropriately attempts to assess the professionalism of the existing social work staff. Professionalism is not the issue at hand. The inclusion of this question does not help to assess the need for a social worker in the emergency room.

The problem with the third question is that there is no way to know how many occasions the respondent might have had a need for a social worker. The emphasis appears to be on availability. Although it is not clear, perhaps the author of this question was trying to explore the time lag between the request for the social worker and the amount of time it took the social worker to disengage from other duties and to appear in the emergency room. If the social workers can always respond within a 5- or 10-minute period, perhaps there is no need to add another social worker just for the emergency room. If this is the case, the evaluator might want to ask the question, "What is the longest you have had to wait for the social worker to arrive in the emergency room?" This question could be followed by another: "About how often does this occur?"

Question 4 is vague and could be improved by asking how often (in terms of times per shift, week, or month) are cases observed where family members need

Example 2 Evaluating In-Service Training

John Practitioner had responsibility for training social workers in a large state agency. John wanted to evaluate a major new training program for supervisors that he firmly believed would make them more effective managers. Knowing how participants at professional training sessions and workshops typically give positive feedback (remember our discussion about client satisfaction studies?), John was determined to go beyond asking, "Did the presenter do a good job?" "Was the presentation clear and well organized?" or "What is your overall rating of this workshop?" Instead, John wanted to know how the week-long workshop would impact trainees as they later performed their jobs. He developed the instrument in Figure 11.2. Will this instrument help him to know the impact the workshop had on the trainees?

specific services like grief counseling or referral to a child protective services agency.

Question 5 seems to repeat the first question. It could be improved by listing a number of situations in which it is likely that emergency room staff would want to have a social worker available to assist. Once again, a frequency count of the times a social worker was needed (e.g., per shift or per week) would supply better information than the vague "occasionally" or "frequently."

See Box 11.6, which discusses the importance of phrasing when determining the questions to be asked.

These questions, drawn from a longer instrument, are straightforward and easy to understand. They do not seem to be vague, double-barreled, leading, and so forth—problems we discussed earlier. There is only one major problem with this collection of items—they measure the respondents' attitudes about whether the training has assisted them. John Practitioner has missed the mark if he is truly interested in the "effect" of the intervention. Despite his best intentions, John has prepared a questionnaire that essentially is just another version of other consumer satisfaction efforts.

BOX 11.6 The Importance of Phrasing

In a study of the reliability of self-reported drug use, sexual behavior, and treatment use, three questions were identified as having high inconsistency rates because they contained poorly defined terms. Evaluators would be well-advised to avoid similar phrasing.

One of these questions asked, "Since your admission, have you had a checkup or have you received any scheduled individual services for medical problems other than those I have already asked about?"

The second question asked, "Since your admission, have you attended any other scheduled talks, lectures, or films as a part of your treatment?"

What is poorly defined in these items? Do you think everyone knows what constitutes "scheduled individual services"? Does a lab test or blood pressure screening qualify, for instance? Similarly, is it possible that some research subjects might not have known when they were participating in a "scheduled talk, lecture, or film"? Could that item be interpreted broadly enough to include lectures in college, or a talk at an art museum?

A third question asked, "Since your admission, how much would you say you have spent on drugs for your own nonmedical use, excluding alcohol?" At least two possibilities for inconsistencies exist here. One explanation could be that the information is unavailable (the respondent has no factual knowledge and is wildly guessing). This would be understandable, because being a drug addict probably precluded the keeping of accurate records. A second possibility is that the drug users were not always purchasing drugs but shared those purchased by someone else and/or exchanged goods and services (e.g., sexual favors) for drugs. And what about over-the-counter drugs? Do they count as drug expenditures?

Source: Adair, E. B. G., Craddock, S. G., Miller, H. G., & Turner, C. F. (1996). Quality of treatment data: Reliability over time of self-reports given by clients in treatment for substance abuse. *Journal of Substance Abuse Treatment,* 13(2), 145–149.

Figure 11.2 Evaluation of the Supervision Workshop

1. Will this training help you to reduce absenteeism among your staff?
   ( ) very little ( ) moderately ( ) very much
2. Will this training help you with the operating costs of your office?
   ( ) very little ( ) moderately ( ) very much
3. Will this training help you to deal with staff's documentation of records?
   ( ) very little ( ) moderately ( ) very much
4. Will this training help you to reduce accident rates among your staff?
   ( ) very little ( ) moderately ( ) very much
5. Will this training help you to increase the productivity of your employees?
   ( ) very little ( ) moderately ( ) very much
6. Will this training help you to get improved ratings from your district manager?
   ( ) very little ( ) moderately ( ) very much

There is nothing wrong with this if that is what the evaluator wanted to accomplish. But in John's case, he wanted to measure the effect or outcomes of the workshop—the transferability to improving functioning and productivity on the job. What John really wanted to measure are such things as:

1. Absenteeism (Is there less absenteeism after the workshop than before?)
2. Operating costs (Are costs lower?)
3. Documentation of records (Are a greater percentage of records in compliance with quality assurance standards?)
4. Accident rates (Are there fewer accidents?)
5. Productivity (Does productivity increase?)
6. Performance ratings (Do performance ratings of supervisors improve?)

In devising his instrument, John opted for a quick measure that examined participants' opinions or attitudes about the workshop. Given his concerns and interests, John would have been better advised to obtain behavioral data such as absenteeism, operating costs, and accident, performance, and productivity rates for the preceding quarter or year for the departmental supervisors' use and to compare those rates with the data after training.

The two examples provided above demonstrate how easy it is to go astray when designing a data collection instrument. Although this chapter furnishes a foundation for understanding what goes into "good" instrumentation, it cannot prepare the reader to anticipate every conceptual problem that may arise or provide the reader with everything he or she needs to know about scale construction. The evaluator who must, by necessity, construct an instrument is always well-advised to have trusted colleagues to review the newly drafted questionnaire or scale and then, after revising it, to pilot test the instrument on a small group of clients or consumers who represent the population to be studied. Such a process will often yield tremendous insights into the way others read and interpret your data collection efforts.

It should also be pointed out that despite our best efforts and sophisticated ways to analyze data and develop scales, they don't always turn out exactly the way we expect or hope. See, for example, the Ferron, Elbogen, Swanson, Swartz, and McHugo (2011) article on assessing the reliability and validity of the Treatment Motivation Questionnaire for people with serious mental illness. The authors point out that the factor structure identified in previous studies did not fit their data and they concluded that more work is needed.

WHAT DO I NEED TO KNOW ABOUT LEVELS OF MEASUREMENT IN DESIGNING INSTRUMENTS?

For most purposes it is not necessary for evaluators to make distinctions between ratio and interval level of measurement when planning how data are to be analyzed. Either level allows for averages that have real meaning to be computed. The data-analytic procedures you'll use to test for differences among groups need ratio or interval level data and are based on the computation of variable means. So creating instruments with summed scores is very acceptable.

The reason statisticians make distinctions between the interval and ratio level of measurements is that a true zero means that comparisons of magnitude have a more precise meaning. In actuality, however, interval and ratio levels of measurement are more alike than they are different. For instance, the Clinical Anxiety Scale (illustrated later in the book) contains 25 items and produces a theoretical range of scores from 0 to 100. The distance between a client scoring 80 at pretest and 40 at posttest can be easily calculated and would represent the same amount of improvement as a client whose pretest was 90 and whose posttest was 50. This is because the intervals between scores are equal, predictable, and form a continuous variable. At some point most variables with many divisions or gradations become interval variables. Variables with as few as 15 or 20 gradations are often treated as interval variables. For all practical purposes, it would not matter to an evaluator if the theoretical range of the Clinical Anxiety Scale was 10 to 90, 20 to 100, or even 25 to 85. The analytical procedures used to analyze the scores will be the same.

Although space limitations prevent further discussion of data analysis in this book, let us conclude this section with some key things to remember about level of measurement:

- For simple description (e.g., 15 percent of the clients were Iraq combat vets), variables measured at the nominal level work fine.
- Ordinal variables also are appropriately used for describing samples ("Twenty percent of the clients slightly improved, 30 percent moderately improved, and 50 percent greatly improved").
- Both nominal and ordinal variables can be used to test whether the proportions in two or more groups are similar.
- However, data recorded at the nominal or ordinal level cannot usually be transformed into interval level data (e.g., if there are two response categories: "under age 60" and "61 or older"—then the data from these two categories will not allow the evaluator to compute the average age of the clients).

- Interval/ratio level of measurement is often desired for dependent variables. If one wants to know whether an intervention significantly reduced the level of depression in a treatment group compared to a control group, then an instrument providing scores on an interval/ratio level would be needed so that group averages could be computed.

Questions for Class Discussion

1. What is wrong with the following questionnaire items?

   a. Describe your mother's condition during her pregnancy with you.
   b. Yes or No: Have you ever been involved in any accidents?
   c. Have you been called names and had your life threatened?

2. Barbara Daydreamer designed a three-item questionnaire to be used as a pre- and posttest instrument to measure adolescents' knowledge of alcoholism as a disease. Later she was surprised to find that there were no significant differences between pre- and posttest scores. How would you explain this?
3. Discuss the following item taken from the evaluation instrument Barbara designed for adolescents: When you are an adult, what are the chances that you will be a drinker?

   _____ I am certain I will never drink
   _____ I don't think I will drink
   _____ I am not sure
   _____ I think I will drink
   _____ I am sure I will drink

4. If you were asked to evaluate an instrument measuring hyperactivity described in a journal, what information about the instrument would you want to find in the article?
5. For a client group you are familiar with, brainstorm what behaviors might be good indicators of treatment success or failure. Why might they work better than measurements of attitudes or knowledge?
6. Suggest one or more variables measured at the interval level that could be converted to nominal or ordinal level variables.

Mini-Projects: Experiencing Evaluation Firsthand

1. Using the Internet or print book resources at your disposal, make a list of at least five scales that you might be able to use in your future practice. Explain how each might be used.
2. Read one of the articles on the development and validation of a scale referenced at the end of this chapter. Summarize, in a short paper, all the steps the author went through.

3. Draft a set of 10 items for a potential scale you would like to see developed. Then, outline a plan to test the scale's reliability and validity. What would you need to do?
4. Develop a needs assessment questionnaire for a program with which you are familiar. Present it to the class for constructive criticism.
5. Group Project: Create a brief questionnaire using only nominal or ordinal variables to measure the sociodemographics of students in your class. Try to come up with approximately 10 items. Have another group construct a similar questionnaire using only variables measured at the interval level. Have a third group examine the data from the two questionnaires and choose the best items from both efforts explaining the reasons for their decisions.

References and Resources

Abell, N., Springer, D., & Kamata, A. (2009). *Developing and validating rapid assessment instruments*. New York: Oxford University Press.

Chao, R. C., & Green, K. E. (2011). Multiculturally sensitive mental health scale (MSMHS): Development, factor analysis, reliability, and validity. *Psychological Assessment*, *23*(4), 876–887.

Ferron, J. C., Elbogen, E. B., Swanson, J. W., Swartz, M. S., & McHugo, G. J. (2011). *Research on Social Work Practice*, *21*(1), 98–105.

Fischer, J., & Corcoran, K. (2013). *Measures for clinical practice and research*. New York: Oxford University Press.

Houck, J. M., Forcehimes, A. A., Gutierrez, E. T., & Bogenschutz, M. P. (2013). Test-retest reliability of self-report measures in a dually diagnosed sample. *Substance Use and Misuse*, *48*(1–2), 99–105.

Johnson, M., Pratt, D. K., Neal, D. B., & Fisher, D. G. (2010). Drug users' test-retest reliability of self-reported alcohol use on the Risk Behavior Assessment. *Substance Use & Misuse*, *45*(6), 925–935.

Koeske, G. F. (1994). Some recommendations for improving measurement validation in social work research. *Journal of Social Service Research*, *18*(3–4), 43–72.

Meisenheimer, C. G. (1985). *Quality assurance: A complete guide to effective programs*. Rockville, MD: Aspen Systems.

Nunnally, J. C., & Bernstein, I. H. P. (1994). *Psychometric theory*. New York: McGraw-Hill.

Ramo, D. E., Hall, S. M., & Prochaska, J. J. (2011). Reliability and validity of self-reported smoking in an anonymous online survey with young adults. *Health Psychology*, *30*(6), 693–701.

Rush, A. J., First, M. B., & Blacker, D. (2007). *Handbook of psychiatric measures*. Washington, DC: American Psychiatric Publishing, Inc.

Sederer, L. I., & Dickey, B. (Eds.). (1996). *Outcomes assessment in clinical practice*. Baltimore: Williams & Wilkins.

Siefert, K., Schwartz, I. M., & Ortega, R. M. (1994). Infant mortality in Michigan's child welfare system. *Social Work*, *39*(5), 574–579.

Sutton, J. E., Bellair, P. E., Kowalski, B. R., Light, R., & Hutcherson, D. T. (2011). Reliability and validity of prisoner self-reports gathered using the Life Event Calendar method. *Journal of Quantitative Criminology*, *27*(2), 151–171.

Thyer, B. A. (1992). Promoting evaluation research in the field of family preservation. In E. S. Morton & R. K. Grigsby (Eds.), *Advancing family preservation practice* (pp. 131–149). Newbury Park, CA: Sage.