6 Evaluation Measures

A Reader’s Guide to Chapter 6

Reliability
  Test-retest, equivalence, homogeneity, and inter- and intrarater reliability

Validity
  Content and face, criterion (predictive and concurrent), and construct validity

Checklist for Creating a New Measure
  Subject matter, item choices, rating scales, format, testing, and revision

Checklist for Selecting an Already Existing Measure
  Costs, content, reliability, validity, and format

The Measurement Chart
  Dependent variables, measures, sampling, timing and duration of measurement, content, reliability, and validity
Reliability and Validity

Evaluators sometimes create their own measures and sometimes adapt or adopt parts or all of already existing ones. Because the conclusions of an evaluation are based on data from the measures used, the quality of the measures must be demonstrably high for the evaluation’s results to be sound. (Otherwise, we have the well-known phenomenon of “garbage in–garbage out.”) To determine the quality of their data collection measures, evaluators must understand the concepts of reliability and validity.

Reliability

A reliable measure is one that is relatively free of “measurement error,” which causes individuals’ obtained scores to be different from their true scores (which can be obtained only through perfect measures). What causes measurement error? In some cases, it results from the measure itself, as when the measure is difficult to understand or poorly administered. For example, a self-administered questionnaire regarding the value of preventive health care might produce unreliable results if it requires a level of reading ability that is beyond that of the teen mothers who are to use it. If the respondents’ reading level is not a problem but the directions are unclear, the measure will also be unreliable. Of course, the evaluator could simplify the measure’s language and clarify the directions and still find measurement error, because such error can also come directly from the people being questioned. For example, if the teen mothers who are asked to complete the questionnaire are especially anxious or fatigued at the time, their obtained scores could differ from their true scores.

In program evaluation, four kinds of reliability are often discussed: test-retest, equivalence, homogeneity, and inter- and intrarater reliability. A measure has test-retest reliability if the correlation between scores on the measure from one time to another is high. Suppose a survey of patient satisfaction is administered in April and again in October to the same group of patients at Hospital A. If the survey is reliable, and no special program or intervention has been introduced between April and October, on average, we would expect satisfaction to remain the same. The major conceptual difficulty in establishing test-retest reliability is in determining how much time is permissible between the first and second administrations. If too much time elapses, external events might influence responses on the second administration; if too little time passes, the respondents may remember and simply repeat their answers from the first administration.

Equivalence, or alternate-form reliability, refers to the extent to which two assessments measure the same concepts at the same level of difficulty. Suppose that students were given an achievement test before participating in a course and then again 2 months after completing it. Unless the evaluator is certain that the two tests were of equal difficulty, better performance on the second test could be a result of that test’s being easier than the first rather than improved learning. And, as with test-retest reliability, because this approach requires two administrations, the evaluator must worry about the appropriate interval between them.
As an alternative to establishing equivalence between two forms of the same instrument, evaluators sometimes compute a *split-half reliability.* This requires dividing an instrument into two equal halves (or alternate forms) and obtaining the correlation between the two halves. Problems arise if the two halves vary in difficulty; however, because only one administration is required, at least the concern over the duration of intervals between testing is eliminated.

*Homogeneity* refers to the extent to which all the items or questions in a measure assess the same skill, characteristic, or quality. Sometimes this type of reliability is referred to as *internal consistency.* The extent of homogeneity is often determined through the calculation of a Cronbach’s coefficient alpha, which is basically the average of all the correlations between each item and the total score. For example, suppose that an evaluator has created a questionnaire to find out about patients’ satisfaction with Hospital A. An analysis of homogeneity will tell the extent to which all items on the questionnaire focus on satisfaction.

Some variables do not have a single dimension. Patient satisfaction, for example, may consist of satisfaction with many elements of the hospital experience: nurses, doctors, financial arrangements, quality of care, quality of surroundings, and so on. If you are unsure of the number of dimensions included in your instrument, you can perform a factor analysis. This statistical procedure identifies relationships among the items or questions in a measure.

*Interrater reliability* refers to the extent to which two or more individuals agree on a given measurement. Suppose that two individuals are sent to a prenatal care clinic to observe patient waiting times, the appearance of the waiting and examination rooms, and the general atmosphere. If these observers agree perfectly in their ratings of all these items, then interrater reliability is perfect. Evaluators can enhance interrater reliability by training data collectors thoroughly, providing them with guidelines for recording their observations, monitoring the quality of the data collection over time to ensure that data collectors are not “burning out,” and offering data collectors opportunities to discuss any difficult issues or problems they encounter in their work.

*Intrarater reliability* refers to a single individual’s consistency of measurement. Evaluators can also enhance this form of reliability through training, monitoring, and continuous education.

**Validity**

*Validity* refers to the degree to which a measure assesses what it purports to measure. For example, a test that asks students to recall information would be considered an invalid measure of their ability to apply information. Similarly, a survey of patient satisfaction cannot be considered valid unless the evaluator can prove that the people who are identified as satisfied on the basis of their responses to the survey think or behave differently than people who are identified as dissatisfied.

*Content validity* refers to the extent to which a measure thoroughly and appropriately assesses the skills or characteristics it is intended to measure. For example, an evaluator who is interested in developing a measure of quality of life for cancer patients has to define *quality of life* and then make sure that items in the measure
adequately include all aspects of the definition. Because of the complexity of this task, evaluators will often consult the literature for models or conceptual frameworks from which to derive the definitions they need. A conceptual model of “quality of life,” for instance, consists of the variables included when the concept is discussed in differing kinds of patients, such as those with cancer, those who are depressed, and those who are very old or very young. It is not uncommon for evaluators to make a statement such as the following in establishing content validity: “We used XYZ cognitive theory to select items on knowledge, and we adapted the ABC Role Model Paradigm for questions about social relations.”

*Face validity* refers to how a measure appears on the surface: Does it seem to ask all the needed questions? Does it use language that is both appropriate and geared to the respondents’ reading level to do so? Face validity, unlike content validity, does not rely on established theory for support.

*Criterion validity* is made up of two subcategories: predictive validity and concurrent validity. *Predictive validity* is the extent to which a measure forecasts future performance. A medical school entry examination that successfully predicts who will do well in medical school has predictive validity. *Concurrent validity* is demonstrated when two assessments agree or a new measure compares favorably with one that is already considered valid. For example, to establish the concurrent validity of a new aptitude test, the evaluator can administer the new measure as well as a validated measure to the same group of examinees and compare the scores. Alternatively, the evaluator can administer the new test to the examinees and compare the scores with experts’ judgment of students’ aptitude. A high correlation between the new test and the criterion measure indicates concurrent validity. Establishing concurrent validity is useful when the evaluator has created a new measure that he or she believes is better (e.g., shorter, cheaper, fairer) than any previously validated measure.

*Construct validity*, which is established experimentally, demonstrates that a measure distinguishes between people who have certain characteristics and those who do not. For example, an evaluator who claims construct validity for a measure of compassionate nursing care will have to prove in a scientific manner that nurses who do well on the measure are more compassionate nurses than nurses who do poorly. Construct validity is commonly established in at least two ways:

1. The evaluator hypothesizes that the new measure correlates with one or more measures of a similar characteristic (convergent validity) and does not correlate with measures of dissimilar characteristics (discriminant validity). For example, an evaluator who is validating a new quality-of-life measure might posit that it is highly correlated with another quality-of-life instrument, a measure of functioning, and a measure of health status. At the same time, the evaluator might hypothesize that the new measure does not correlate with selected measures of social desirability (individuals’ tendency to answer questions so as to present themselves in a positive light) or measures of hostility.

2. The evaluator hypothesizes that the measure can distinguish one group from the other on some important variable. For example, a measure of compassion should be able to demonstrate that people who are high scorers are compassionate and that people who are low scorers are unfeeling. This
requires that the evaluator translate a theory of compassionate behavior into measurable terms, identify people who are compassionate and people who are unfeeling (according to the theory), and prove that the measure consistently and correctly distinguishes between the two groups.

A Note on Language: Data Collection Terms

The language used to discuss reliability and validity (terms such as examinees, scores, scales, tests, and measures) comes from test theory, or psychometrics. Program evaluators often use the terms data source, measure, scale, test, and instrument interchangeably. This is sometimes confusing, to say the least, especially because evaluators also talk about outcome measures and outcome indicators when referring to evaluation study outcomes. The following brief lexicon can be helpful for sorting out data collection terms.

A Guide to Data Collection Terms

- **Data source**: Any source of information for the evaluation. This may include data from questionnaires or tests, literature reviews, existing databases (such as Medicare’s), and vital statistics (such as the number of live births in a given year).
- **Index**: A way of rank ordering things. Scores on an index of function give an indication of where people stand in relation to one another. This term is sometimes used interchangeably with scale.
- **Instrument**: A device or strategy used to collect data; instruments include laboratory tests, self-administered questionnaires, and interviews. This term is often used interchangeably with measure.
- **Measure**: Very much like a data source. This term is often used interchangeably with instrument, test, and assessment.
- **Outcome**: The consequences of participating in a program. Outcomes may be changes in areas such as health status and emotional well-being.
- **Outcome measure or outcome indicator**: Often used as a synonym for outcome.
- **Rating scale**: A graded set of choices. Scales may be nominal (or categorical) (e.g., race, gender), ordered (or ordinal) (e.g., with response categories such as often, sometimes, and never; Stages I, II, and III of a disease), or numerical, including continuous (age, height) and discrete (number of pregnancies, number of arrests for driving drunk). The most commonly used type of rating scale is the Likert scale, with response categories such as strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree.
- **Scale**: A combination of items or questions that measure the same concept, such as a 10-item scale that measures emotional well-being or a 36-item scale that measures health status.
- **Test**: Achievement test, laboratory test.
Checklist for Creating a New Measure

Knowing the types of measures that are available and how to demonstrate their reliability and validity enables the evaluator to get down to the serious business of developing a measure that is tailored to the needs of the investigation or selecting and adapting one that is already in use. Before you attempt to create a new measure for your evaluation study, you must make sure that you have identified the domain of content (through observation or with the help of experts, research, and theory) and have the expertise, time, and money to complete the task. The following is a checklist of the basic steps you need to take in creating a new measure.

1. Set boundaries.
   - Decide on the type of measure (e.g., questionnaire, observation).
   - Determine the amount of needed and available time for administration and scoring (e.g., a 15-minute interview and 10 minutes for summarizing responses).
   - Select the kinds of reliability and validity information to be collected (e.g., to establish alternate-form reliability, you must develop two forms; to establish concurrent validity, you will need an already existing instrument).

2. Define the subject matter or topics that will be covered. For definitions, consult the literature, experts, or health care consumers. Example 6.1 illustrates how the definitions for an evaluation of prenatal care programs were found in the literature and corroborated by nurses, physicians, and evaluators.

Example 6.1 Defining Terms: The Case of Prenatal Care Programs

Prenatal health care refers to pregnancy-related services provided to a woman between the time of conception and delivery and consists of monitoring the health status of the woman; providing patient information to foster optimal health, good dietary habits, and proper hygiene; and providing appropriate psychological and social support. Programs have preset, specific purposes and activities for defined populations and groups. Outcomes of prenatal care programs include the newborn’s gestational age and birth weight and the mother’s medical condition and health habits.

These definitions tell the evaluator something like the following: If you want to evaluate prenatal care programs, your measures should include attention to patient education, dietary habits and hygiene, and psychosocial support. If you are interested in outcomes, you will need measures of gestational age and mother’s medical condition and health habits. You will also need to decide which medical conditions (e.g., diabetes, hypertension) and health habits (e.g., drinking, smoking) will be the focus of your evaluation.
3. Outline the content.

Suppose that an evaluator is concerned with the outcomes of a particular prenatal care program: the Prenatal Care Access and Utilization Initiative. Assume also that the evaluator’s review of the literature and consultation with experts reveal the importance of collecting data on the following variables: still or live birth, birth weight, gestational age, number of prenatal visits, and drug toxicology status of mother and baby. An outline of the contents might look like this:

I. Baby’s birth date
II. Birth weight
III. Gender
IV. Gestational age
V. Whether a drug toxicology screen was performed on baby and results
VI. Whether a drug toxicology screen was performed on mother and results
VII. Number of visits

4. Select item choices.

An item on a measure is a question asked of the respondent or a statement to which the respondent is asked to react. Example 6.2 presents a sample item and its answer choices.

Example 6.2  Item Choices

<table>
<thead>
<tr>
<th>Item: What are the test results?</th>
<th>Choose one</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td></td>
</tr>
<tr>
<td>Normal or negative ................</td>
<td>☐</td>
</tr>
<tr>
<td>Not significant but abnormal ......</td>
<td>☐</td>
</tr>
<tr>
<td>Positive ..........................</td>
<td>☐</td>
</tr>
<tr>
<td>Equivocal/No Data ..................</td>
<td>☐</td>
</tr>
</tbody>
</table>

Selecting choices for items requires skill and practice. Whenever possible, you should use item choices that others have used effectively. The possibilities of finding appropriate choices are greater when you are collecting demographic information.
(e.g., age, gender, ethnicity, income, education, where a person lives), for example, than when you are collecting data on the knowledge, attitudes, or behaviors that result from a specific program designed for a particular group of people. You can find effective item choices in the literature and can obtain many from measures prepared by U.S. Bureau of the Census; the health departments of cities, counties, and states; and other public and private agencies.

5. Choose rating scales.

You should adapt rating scales from scales that have already been proven in earlier research whenever possible. Like the choices for items, rating scales are available from measures designed by public and private agencies and those described in the literature. Example 6.3 displays an item that uses a simple true-and-false scale.

### Example 6.3 Item With a True-False Rating Scale

Please circle the number that best describes whether each of the following statements is true or false for you.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Definitely True</th>
<th>Mostly True</th>
<th>Not Sure</th>
<th>Mostly False</th>
<th>Definitely False</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am somewhat ill.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I am as healthy as anybody I know.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>My health is excellent.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I have been feeling bad lately.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

6. Review the measure with experts and potential users.

It is wise to ask other evaluators, subject-matter experts, and potential users to review your measure. The following are some important questions to ask them.

### Questions to Ask of Those Reviewing Your Measure

- **Ask experts**
  1. Is all relevant content covered?
  2. Is the content covered in adequate depth?
3. Are all item choices appropriate?
4. Are all rating scales appropriate?
5. Is the measure too long?

• Ask users
  1. Is all relevant content covered?
  2. Is the content covered in adequate depth?
  3. Do you understand without ambiguity all item choices and scales?
  4. Did you have enough time to complete the measure?
  5. Did you have enough time to administer the measure?
  6. Is the measure too long?

7. Revise the measure based on comments from the reviewers.

8. Put the measure in an appropriate format. For example:
   - Add an ID code, because without such coding you cannot collect data on the same person over time.
   - Add directions for administration and completion.
   - Add a statement regarding confidentiality (informing the respondent that participants are identifiable by code) or anonymity (informing the respondent that you have no means of identifying participants).
   - Add a statement thanking the respondent.
   - Give instructions for submitting the completed measure. If it is to be mailed, is an addressed and stamped envelope provided? By what date should the measure be completed?

9. Review and test the measure before administration.

The importance of pilot-testing a new measure cannot be overemphasized. To conduct a meaningful pilot test, you must use the measure under realistic conditions. This means administering the measure to as many participants as your resources will allow. After they have completed the measure, you need to interview the participants to find out about any problems they might have had in completing the measure. When your study involves interviews, you must test the methods for interviewing as well as the measure itself.

Checklist for Selecting an Already Existing Measure

Many instruments and measures are available for use by health program evaluators. Good sources for these are the published evaluation reports found in journals. In some cases, whole instruments are published as part of these articles. Even when the instruments themselves do not appear, the evaluators usually describe all of their main data sources and measures in the “methods” sections of their reports, and you can contact them for any additional information you need.
Using an already tested measure has many advantages, among which is that it saves you the time and other resources needed to develop and validate a completely new instrument. Choosing a measure that has been used elsewhere is not without pitfalls, however. For example, you may have to pay to use an established measure, you may be required to share your data, and you may have to modify the measure so substantially that its reliability and validity are jeopardized, requiring you to establish them all over again.

The following is a checklist for choosing an already existing measure.

1. **Find out the costs:** Do you have to pay? Share data? Share authorship?

2. **Check the content:** In essence, you must do your own fact and content validity study. Make sure that the questions are the ones you would ask if you were developing the instrument. Check the item choices and rating scales. Will they get you the information you need?

3. **Check the reliability and validity:** Make sure that the types of reliability and validity that have been confirmed are appropriate for your needs. For example, if you are interested in interrater reliability but only internal consistency statistics are provided, the measure may not be the right one for you. If you are interested in a measure’s ability to predict but only content validity data are available, think again before adopting the instrument.

You will need to check carefully the context in which the measure has been validated. Are the settings and groups similar to those in your evaluation? If not, the instrument may not be valid for your purposes. For example, a measure of compliance with counselors’ advice in a program to prevent child abuse and neglect that has been tested on teen mothers in Montana may not be applicable to nonteen mothers in Helena, Montana, or to teen mothers in San Francisco, California.

You must also decide whether the measure is sufficiently reliable and valid for use. Reliability and validity are often described as correlations (say, between experts or measures or among items). How high should the correlations be? The fast answer is that the higher, the better, and .90 is best. But the statistic by itself should not be the only or even the most important criterion. A lower correlation may be acceptable if the measure has other properties that are potentially more important. For example, the content may be especially appropriate or the measure might have been tested on participants who are very much like those in your evaluation.

4. **Check the measure’s format:**
   - Will the data collectors be able to score the measure?
   - Does it make sense to use it, given your available technology? For example, if it requires certain software or expertise, do you have it? Can you afford to get it?
   - Will the participants in the evaluation be willing to complete the measure? Participants sometimes object to spending more than 10 or 15 minutes on an interview, for example. Also, personal questions and complicated instructions can result in incomplete data.
The Measurement Chart: Logical Connections

A measurement chart assists the evaluator in the logistics of the evaluation by helping him or her to be sure that all variables will have the appropriate coverage. Such a chart is also useful when the evaluator is writing proposals, because it portrays the logical connections among what is being measured, how, for how long, and with whom. When the evaluator is writing reports, the chart provides a summary of some of the important features of the evaluation’s data sources. As the sample measurement chart in Figure 6.1 illustrates, the information in the chart’s columns enables the evaluator to make logical connections among the various segments of data collection. Each column in the chart is explained briefly below.

Variables. To ensure that all independent and dependent variables are covered, the evaluator uses the chart to check the evaluation questions and sampling strategies, including all strata and inclusion and exclusion criteria. For example, suppose that an evaluation asks about the effectiveness of a yearlong combined diet and exercise program in improving the health status and quality of life for people over 75 years of age. Suppose also that it excludes all persons with certain diseases, such as metastatic cancer and heart disease. Assume that the evaluators plan to compare men and women to determine whether any differences exist between them after program participation. The variables needing measurement in such an evaluation would include quality of life, health status (to identify persons with metastatic cancer and heart disease and to assess changes), and demographic characteristics (to determine who is male and who is female).

How measured. For each variable, the measure should be indicated. The measurement chart in Figure 6.1 shows that quality of life will be assessed through interviews with patients and observations of how they live, health status will be measured through physical examination, demographic characteristics will be measured through self-administered questionnaires or interviews, and costs will be measured through a review of financial records.

Sample. This column in the measurement chart contains information on the number and characteristics of individuals who will constitute the sample for each measure. For example, the measurement chart in Figure 6.1 shows that to measure quality of life the evaluator will interview all 100 patients (50 men and 50 women) in the experimental group and all 100 patients (50 men and 50 women) in the control group as well as observe a sample of the lifestyles of 50 patients (25 men and 25 women) in each group. Assessment of health status will be based on physical examination of all persons in the experimental and control groups, and demographic information will be collected on all experimental and control program participants. Data on costs will be collected only for those individuals who use one of the two staffing models.
<table>
<thead>
<tr>
<th>Dependent Variables</th>
<th>How Measured</th>
<th>Sample</th>
<th>Timing of Measures</th>
<th>Duration of Measures</th>
<th>Content</th>
<th>Reliability and Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality of life</td>
<td>Interviews with patients</td>
<td>All 100 patients in experimental group and all 100 in control. In each group of 100, 50 men and 50 women</td>
<td>1 month before program participation, 1 month after, and 1 year after</td>
<td>1-hour interviews; 30 minutes to summarize</td>
<td>Social, emotional, and physical functioning; health beliefs; perceived joy and satisfaction: 35 questions</td>
<td>The Brandyse Functional Assessment (long form) and the University Quality of Living Scale will be adopted. The Brandyse has been validated on 4,000 community-dwelling elderly. Test-retest reliability is .85 and homogeneity for subscales is .90. No costs will be incurred; data will be shared with the Brandyse group. The University Quality of Living Scale is available free to all researchers. All patients will complete informed consent forms. No special institutional review board-type procedure need be followed. No special software or hardware is necessary.</td>
</tr>
<tr>
<td>Observations</td>
<td>50 patients in the experimental and 50 in the control groups; 25 men and 25 women in each</td>
<td>1 month before program participation and 6 months after</td>
<td>Half-hour observations; 15 minutes to summarize</td>
<td>Appearance and repair of household; number of visitors; fire and accident safety: 10 questions</td>
<td>Interrater reliability will be estimated between at least two observers.</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6.1.** Measurement Chart for an Evaluation of a Health Care Program's Effects on Patient Quality of Life and Health Status (Continued)
<table>
<thead>
<tr>
<th>Dependent Variables</th>
<th>How Measured</th>
<th>Sample</th>
<th>Timing of Measures</th>
<th>Duration of Measures</th>
<th>Content</th>
<th>Reliability and Validity</th>
<th>General Concerns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health status</td>
<td>Physical examination</td>
<td>All persons in experimental and control groups</td>
<td>1 month before the program, within 1 month of completion, and 1 year later</td>
<td>30 minutes</td>
<td>Emphasis on presence or absence of serious chronic disorders (e.g., metastatic cancer, heart disease): 50 questions</td>
<td>A team of four physicians and nurse practitioners will be trained to administer the physical examinations in a uniform way.</td>
<td></td>
</tr>
<tr>
<td>Demographic characteristics</td>
<td>Self-administered questionnaires or interviews</td>
<td>All persons in experimental and control groups</td>
<td>1 month before the start of the program</td>
<td>Less than 5 minutes</td>
<td>Gender, ethnicity, age, education, household income, region of country in which highest level of education was achieved: 10 questions</td>
<td>Standard items will be used to collect these data.</td>
<td></td>
</tr>
<tr>
<td>Costs</td>
<td>Financial records</td>
<td>All persons receiving care in two clinics, one staffed primarily by physicians and one by nurses</td>
<td>Within 1 month of completion of the program</td>
<td>About 30 minutes to obtain data and make calculations</td>
<td>Number of staff; hourly wages; number of appointments to each clinic made and kept; number of minutes spent on care</td>
<td>A form will be created and data collectors will be trained to use it. Periodic quality checks will be made.</td>
<td></td>
</tr>
</tbody>
</table>
Timing of measures. The information in this column refers to when each measure is to be administered. For example, the measurement chart in Figure 6.1 shows that interviews regarding quality of life and physical examination will be conducted 1 month before the program, immediately after the program (within 1 months), and 1 year after. Observations will be made 1 month before and 6 months after. Demographic information will be obtained just once: 1 month before the start of the program.

Duration of measures. This column of the chart contains information on the amount of time it will take to administer and summarize or score each measure. The measurement chart in Figure 6.1 shows that the quality-of-life interviews will take 1 hour to conduct and a half hour to summarize. The observations will take a half hour to conduct and 15 minutes to summarize. The physical examinations are expected to take 30 minutes, and collecting data on demographic characteristics will take less than 5 minutes.

Content. The evaluator should provide a brief description of the content in the measurement chart. For example, if measurement of quality of life is to be based on a particular theory, the evaluator should note the theory’s name. If the interview has several sections (e.g., social, emotional, and physical function), the evaluator might mention them. It is important to remember that the chart’s purpose is really to serve as a guide to the measurement features of an evaluation. Each one of its sections may require elaboration. For example, for some measures, the evaluator may want to include the number of items in each subscale.

Reliability and validity. If the measures being used are adapted from some other study, the evaluator might describe the relevant types of reliability and validity statistics in this part of the chart. For example, if the quality-of-life measure has been used on elderly people in another evaluation that showed that higher scorers had higher quality than low scorers, this information might be included. If additional reliability information is to be collected in the current evaluation, that, too, may be reported. A review of medical records to gather information on the number, types, and appropriateness of admissions to the hospital over a 1-year period, for example, could require estimations of data collectors’ interrater reliability; such information belongs in this section of the chart.

General concerns. In this portion of the chart, the evaluator notes any special features of the entire data collection and measurement endeavor. These include information on costs, training, number of items, special software or hardware requirements, and issues pertaining to informed consent.

Summary and Transition to the Next Chapter on Managing Evaluation Data

Reliability refers to the consistency of a measure, and validity refers to its accuracy. Having reliable and valid measures is essential in a diligent evaluation. Sometimes the
evaluator is required or chooses to create a new measure; at other times, a measure is available that appears to be suitable. Whether creating, adapting, or adopting, the evaluator must critically review the measure to ensure its appropriateness and accuracy for the current study.

A measurement chart is a useful way of showing the relationships among dependent variables, how and when the dependent variables are measured, and the content, reliability, and validity of the measures. Measurement charts are useful tools for evaluators as they plan and report on evaluations.

The next chapter discusses the activities in which the evaluator engages to ensure the proper management of evaluation data to preserve them for analysis. These activities include drafting an analysis plan, creating a codebook, establishing coder reliability, reviewing data collection instruments for incomplete or missing data, entering data into a database, cleaning the data, creating the final data set, and archiving the data set.
Exercises

EXERCISE 1: RELIABILITY AND VALIDITY

Directions

Read the following excerpts and determine which concepts of reliability and validity are covered in each.

Excerpt A

The self-administered questionnaire was adapted with minor revisions from the Student Health Risk Questionnaire, which is designed to investigate knowledge, attitudes, behaviors, and various other cognitive variables regarding HIV and AIDS among high school students. Four behavior scales measured sexual activity (4 questions in each scale) and needle use (5 questions); 23 items determined a scale of factual knowledge regarding AIDS. Cognitive variables derived from the health belief model and social learning theory were employed to examine personal beliefs and social norms (12 questions).

Excerpt B

All charts were reviewed by a single reviewer with expertise in this area; a subset of 35 charts was reviewed by a second blinded expert to assess the validity of the review. Rates of agreement for single items ranged from 81% ($\kappa = .77; p < .001$) to 100% ($\kappa = 1; p < .001$).

Excerpt C

Group A and Group B nurses were given a 22-question quiz testing evaluation principles derived from the UCLA guidelines. The quizzes were not scored in a blinded manner, but each test was scored twice.

EXERCISE 2: REVIEWING A DATA COLLECTION PLAN

Directions

Read the following information collection scenario and, acting as an independent reviewer, provide the evaluator with a description of your problems and concerns.

The School of Nursing is in the process of revising its elective course in research methods. As part of the process, a survey was sent to all faculty who currently teach the methods courses to find out whether and to what extent epidemiology topics were included. Among the expectations was that methods courses would aim to improve students’ knowledge of epidemiology and their attitudes toward its usefulness in a number of nursing subjects, ranging from public health nursing to home health care administration. The results
of the survey revealed little coverage of some important objectives. Many faculty indicated that they would like to include more epidemiology, but they were lacking educational materials and did not have the resources to prepare their own. To rectify this, a course with materials was developed and disseminated. The Center for Nursing, Education, and Evaluation was asked to appraise the effectiveness of the educational materials. Evaluators from the center prepared a series of knowledge and skill tests and planned to administer them each year over a 5-year period. The evaluators are experts in test construction, and so they decided to omit pilot testing and save the time and expense. Their purpose in testing was to measure changes (if any) in nurses’ abilities. They also planned to interview a sample of cooperating students to get an in-depth portrait of their knowledge of clinical epidemiology.
Suggested Readings


This article describes the validation of the Cancer Rehabilitation Evaluation System (CARES). It provides an excellent example of the way in which a new measure is tested.


This book is an excellent source of information on the different types of reliability and validity.


This book provides numerous examples of health measurement techniques and scales and also discusses the validity and reliability of important health measures.


This article describes the statistics used to establish the reliability and validity of a widely used measure of health.


This article provides definitions and examples of the concepts of reliability and validity.

*Note: More information on reliability and validity is available on the Internet. Use your favorite search engine and the search phrase “reliability and validity.”*
Purpose of This Chapter

This chapter provides information on how to organize and manage the data collected during an evaluation so that it can be analyzed. The topics addressed include the kinds of software available to assist evaluators in data entry and storage, how to monitor data entry to avoid errors, and how to create a “clean” data set. Because evaluators often collect a great deal of data, even in relatively small evaluations, it is very important that they manage their data efficiently.