Designing Questions

When social scientists measure variables by having people respond to questions or statements, this can be done in one of two ways: with *questionnaires* or with *interviews*. A **questionnaire** contains written questions that people respond to directly on the questionnaire form itself, without the assistance of an interviewer. An **interview** involves an interviewer reading questions to respondents and recording their answers. It is important to keep in mind that when social scientists talk about their "measuring instrument," they are referring to the whole questionnaire or interview, not just to the individual questions. After all, people are affected by and react to the whole thing, including such aspects as the physical appearance and the order in which questions or statements are presented. Chapter 9 explores how to develop whole questionnaires and interviews, including such aspects as physical appearance and question ordering. Chapter 6 focuses on a more limited issue: how to design the individual questions, statements, indexes, and scales that make up questionnaires and interviews. Designing valid questions and statements is a complex and challenging process, and I will review some of the major guidelines that can be used in doing it well (Sheatsley, 1983).

Conceptual Relevance

To design good questions, a researcher needs to be clear about the nature and dimensions of the concepts to be measured and take care that the questions link directly with the concepts. For example, if the concepts relate to people's behavior, then the questions should ask about behaviors—not attitudes, thoughts, or emotions. Likewise, concepts that refer to emotions or attitudes should be tapped by questions that address emotions or attitudes. As another example, a concept that relates to poverty is not necessarily adequately measured by questions having to do with welfare. Although poverty and welfare are related, they are not identical, and researchers need to continually relate the questions being developed to the abstract concepts they are intended to measure. This again illustrates the critical importance of conceptual development discussed in Chapter 2 and Chapter 4 and the continual interplay between the abstract conceptual level and the concrete level of operational definitions.

Directions

One of the simplest but most important parts of a measuring device is the directions that guide the respondents' answering of the questions. Good directions go a long way toward improving the quality of data generated by questionnaires and interviews. If you want respondents to put an X in a box corresponding to their answer, tell them precisely that. Questionnaires often contain items requiring different ways of answering. At each place in the questionnaire where the format changes, additional directions should be included. As the development of questions, indexes, and scales are described in this chapter, examples of how to design good directions will be provided. Clear directions are especially important in questionnaires since there may be no one present to clarify any ambiguities, as when a questionnaire is distributed through the mail or over the Internet. However, directions are also important in interviews where the interviewer may read directions that guide the respondent's answers.

Closed-Ended versus Open-Ended Questions

Two basic types of questions can be used: *closed ended* or *open ended* (Sudman & Bradburn, 1982). **Closed-ended questions** are those that provide respondents with a fixed set of alternatives from which to choose. The response formats of multiple-item scales, for example, are all closed ended, as are multiple-choice examination questions, with which you are undoubtedly familiar. **Open-ended questions** are questions to which the respondents write their own responses, much as you do for an essay-type examination question.

The proper use of open- and closed-ended questions is important for the quality of data generated. Theoretical considerations play an important part in the decision about which type of question to use. In general, closed-ended questions should be used when all the possible, theoretically relevant responses to a question can be determined in advance and the number of possible responses is limited. For example, a question relating to marital status would almost certainly be treated as a closed-ended question. A known and limited number of answers are possible: married, single, divorced, separated, or widowed. In current research, people may be offered an additional alternative—"living together" or "cohabitating"—in order to reflect more accurately the living arrangements that people choose today. Another obvious closed-ended question is about sexual status. To leave such questions open ended runs the risk that some respondent will either purposefully or inadvertently answer in a way that provides meaningless data. Putting "sex" with a blank after it, for example, is an open invitation for some character to write "yes" rather than give the information needed.

Open-ended questions, on the other hand, are used in qualitative research when the researcher wants individuals to describe their feelings or discover the meanings that are important to people. For such data, open-ended questions allow people to provide in their own words as complete an accounting as possible of some phenomenon. Data of this sort will be discussed in more detail in Chapter 11 and Chapter 15. Quantitative research also sometimes uses open-ended questions, but in this case the ultimate goal will probably be to quantify people's responses. Open-ended questions might therefore be appropriate in an exploratory study in which the lack of theoretical development makes it difficult to know how to categorize people's responses before seeing the responses. In addition, when researchers cannot predict all the possible answers to a question or when too many possible answers exist to list them all practically, then open-ended questions are appropriate. Suppose we wanted to know the reasons people moved to their current residence. So many possible reasons exist that such a question would probably be treated as open ended. If interested in the county and state in which respondents reside, we could generate a complete list of all the possibilities and thus create a closed-ended question. But the list would consume so much space on the questionnaire that it would be excessively cumbersome, especially considering that respondents should be able to answer this question correctly in its open-ended form.

Some questions lend themselves to a combination of both formats. Religious affiliation is a question usually handled in this way. Although there are a great many religions, there are some to which only a few respondents will belong. Thus, religions with large memberships can be listed in closed-ended fashion with a category "Other," where a person can fill in the name of a religion not found on the list (see question 4, Table 6.1). Any question with a similar pattern of responses—numerous possibilities, but a few popular ones—can be handled efficiently in this way. The combined format maintains the convenience of closed-ended questions for most of the respondents but also allows those with less common responses to express them. When the option of "Other" is used in a closed-ended question, it is a good idea to prompt respondents to write in their response by indicating "please specify." These answers can then be coded into whatever response categories seem appropriate for data analysis.

Another issue in choosing between open- and closed-ended questions is the ease with which each can be handled at the data-analysis stage. Closed-ended questions can be quickly entered into a computer file and prepared for data analysis. Open-ended questions, if they are to be handled quantitatively, must first be coded, which means creating a category system into which everybody's answers can be placed. The process of coding is discussed in chapters 11, 13, and 15, but the point here is that it is a more time-consuming form of data analysis. If the nature of the variables warrants using closed-ended questions, it will be quicker and more efficient at the data-analysis stage.

As you can see from this discussion, the decision about whether to use open- or closed-ended questions is complex and can have substantial effects on the type and quality of the data that are collected, as was illustrated in a survey of attitudes about social problems confronting the United States. The Institute for Social Research at the University of Michigan asked a sample of respondents open- and closed-ended versions of essentially the same questions (Schuman & Presser, 1979). The two versions elicited quite different responses. For example, with the closed-ended version, 35% of the respondents indicated that crime and violence were important social problems, compared to only 15.7% in the open-ended version. The same pattern occurred with a number of other issues: People were more likely to identify an issue as a problem if it was on the closed-ended list. One reason that the type of question has such an effect is that the list of alternatives in the closed-ended questions tends to serve as a "reminder" to the respondent of issues that might be problems. Without the stimulus of the list, some respondents might not even think of those issues. A second reason is that people tend to choose from the list provided in closed-ended questions rather than writing in their own answers even when provided with an "Other" category.

It is possible, in some cases, to gain the benefits of both open- and closed-ended questions by using an open-ended format in a pretest or pilot study and then, based on the

TABLE 6.1 Formatting Questions for a Questionnaire

Please indicate your response to the following questions by placing a X in the appropriate box.

1. Which of the following best describes where you live?

   ☒ In a large city (100,000 population or more).

   □ In a suburb near a large city

   □ In a middle-sized city or small town (under 100,000 population) but not a suburb of a large city.

   □ Open country (but not on a farm)

   □ On a farm

2. Have you ever shoplifted an item with a value of $10 or more?

   □ Yes

   □ No

   If Yes: How many times have you taken such items?

   □ Once

   □ 2 to 5 times

   □ 6 to 10 times

   □ More than 10 times

Filter Questions

Contingency Question

3. Have you purchased a new automobile between 1995 and the present?

   □ Yes

   □ No (*If No,* please skip to Section C, question 1.)

4. Please indicate the religion to which you belong:

   □ Protestant

   □ Catholic

   □ Jewish

   □ Other. Please specify. ____________________

results, designing closed-ended questions for the actual survey.

Wording of Questions and Statements

Because the questions or statements that are used to measure variables are the basic data-gathering devices, they need to be worded with great care. This is true especially for questionnaires that allow no opportunity to clarify questions for the respondent. With these questionnaires, ambiguity in questions can be a source of substantial measurement error. I will review some of the major issues in developing good questions (Gorden, 1992; Sudman & Bradburn, 1982).

Pretesting

The wording of questions should always be subjected to empirical assessment. In other words, whenever possible, a researcher should attempt to determine whether a particular wording might lead to unnoticed bias. Words, after all, have connotative meanings (that is, emotional or evaluative associations) that the researcher may not be aware of but that may influence respondents' answers to questions. In a study of people's attitudes about social welfare policy in the United States, for example, survey respondents were asked whether they believed we should spend more or less money on welfare (T. W. Smith, 1987). However, they were asked the question in three slightly different ways: Group one was asked whether we were spending too much or too little on "welfare"; group two was asked about spending on "assistance for the poor"; and group three was asked about money for "caring for the poor." At first glance, all three questions would seem to have much the same meaning; yet people's responses to them suggested something quite different. Basically, people responded far more negatively to the question containing the word *welfare*, indicating substantially less willingness to spend more money on "welfare" than they were to "assist the poor." This seems to be a minor semantical difference in wording, yet the impact was dramatic. Although the study didn't investigate why these differing responses occurred, it seems plausible that, for many people, the word *welfare* has connotative meanings that involve images of laziness, waste, fraud, bureaucracy, or the poor as disreputable. "Assisting the poor," on the other hand, is more likely to be associated with giving and Judeo-Christian charity. These connotations lead to very different responses. In many cases, the only way to assess such differences is to compare people's responses to different versions of the same question during a pretest or to conduct assessments of validity and reliability, as discussed in Chapter 5.

So, once questions and statements are developed, they should be pretested to see if they are clearly and properly understood and are unbiased. Pretesting can be done by having people respond to the questionnaire or interview and then reviewing their responses with them to find any problems. The way in which a group responds to the questions themselves can also point to trouble. For example, if many respondents leave a particular answer blank, then there may be a problem with the question. Once the instrument is pretested, modifications should be made where called for, and it should be pretested again. Any change in the questionnaire requires more pretesting. Once it is pretested with no changes called for, it is ready to be used in research.

Tense

In general, questions should be stated in the present tense. An exception would be specialized questions that focus on past experiences or expectations for the future. In these situations, the appropriate tense would be used. Of major importance is that tenses not be mixed carelessly. Failure to maintain consistent tense of questions can lead to an understandable confusion on the part of respondents and therefore more measurement error.

Simple, Direct, and Clear

Questions should be simple, direct, and express only one idea. Complex statements expressing more than one idea should be avoided. For example, the first statement illustrated in Table 6.2 is not acceptable for this reason and needs to be changed as indicated. Another practice to avoid is reference to things that are not clearly definable or that depend on the respondent's interpretation. This is a problem with the last two questions in Table 6.2, and a revision is suggested as a solution.

Always be aware that statements that seem crystal clear to a researcher may prove unclear to many respondents. When writing questions, it is preferable to err on the side of designing questions that may be too simple for the intended audience than to err in the opposite direction. This concern is especially relevant when respondents are suspected of having low levels of education or poor reading skills. Accordingly, the researcher should avoid the use of technical jargon. For example, it would not be advisable to include a statement that reads, "The current stratification system in the United States is too rigid." *Stratification* is a technical

TABLE 6.2 Common Errors in Writing Questions and Statements

Original Question	Problem	Solution
The city needs more housing for the elderly and property taxes should be raised to finance it	**Two questions in one:** Some respondents might agree with the first part but disagree with the second.	Questions should be broken up into two separate statements, each expressing a single idea.
In order to build more stealth bombers, the government should raise taxes.	**False premise:** What if a person doesn't want more bombers built? How do they answer?	First ask their opinion on whether the bomber should be built; then, for those who respond "Yes," ask the question about taxes.
Are you generally satisfied with your job, or are there some things about it that you don't like?	**Overlapping alternatives:** A person might want to answer "Yes" to the first part (i.e., they are generally satisfied) but "No" to the second part (i.e., there are also some things they don't like).	Divide this into two questions: one measures their level of satisfaction while the other assesses whether there are things they don't like.
How satisfied are you with the number and fairness of the tests in this course?	**Double-barreled question:** It asks about both the "number" and the "fairness," and a person might feel differently about each.	Divide this into two questions.
What is your income?	**Vague and ambiguous words:** Does "income" refer to before-tax or after-tax income? To hourly, weekly, monthly, or yearly income?	Clarify: What was your total annual income, before taxes, for the year 2000?
Children who get into trouble typically have had a bad home life.	**Vague and ambiguous words:** The words *trouble* and *bad home life* are unclear. Is it trouble with the law, trouble at school, trouble with parents, or what? What constitutes a *bad home life* depends on the respondent's interpretation.	Clarify: Specify what you mean by the words: *trouble* means "having been arrested" and *bad home life* means "an alcoholic parent."

word used in the social sciences that many people outside the field may not understand in the same sense that social scientists do.

Length

Questions and statements should be kept as short as possible while still providing the respondent with the complete and accurate intended meaning. One guideline that has been offered is that questions not be more than 25 words in length (Sheatsley, 1983). While this is not a hard and fast rule, it would certainly be wise to look critically at any question approaching 25 words to determine whether it is potentially confusing or actually contains more than one thought and should be broken down into two or more shorter questions.

Insider Language and Slang

For most questions, especially those designed for the general public, slang words should not be used. Slang usage tends to arise in the context of particular groups and subcultures. Slang words may have a precise meaning within those groups but confuse people outside those groups. Occasionally, however, the target population for a survey will be more specialized than the general population, and the use of their insider language or "in-group" jargon may be appropriate. It would demonstrate to the respondents that the researcher cared enough to "learn their language" and could increase rapport, resulting in better responses. In interviewing prostitutes, for example, it might be preferable to refer to prostitutes as "girls" or "working girls" (if this is the terminology used by the prostitutes themselves); referring to them as "prostitutes" may create social distance and make it more difficult to gather valid responses from them. Having decided to use slang, however, the burden is on the researcher to be certain the slang is used correctly.

Some additional problems that can arise in writing questions and statements are presented in Table 6.2.

Providing a Context

To elicit valid responses, it is important to be sure that people possess adequate information to give accurate answers. In part, this means asking specific questions, but it also means placing the questiôn in a clear context for the respondent. One way to provide a clear context is to place the question and answer in a specific *time perspective*. For example, if you were interested in individuals' perceptions of changes in their financial circumstances, you might ask "Do you think that your financial circumstances are getting better or worse?" Yet, two people whose circumstances are identical might give very different responses if one interpreted the question to mean "over the past few months" while the other was looking "over the past few years." Based on whatever time period seems appropriate given conceptual considerations, you could revise the question to read, "Considering just the past two-year period, do you think that your financial circumstances are getting better or worse?"

Another way to provide a context for a question is to provide a *spatial* or *geographic perspective*. A quality of life study, for example, might be interested in whether people feel safe in their community. You might ask, "Do you feel safe walking the streets at night alone?" but this could elicit different responses if the person is thinking of his or her own neighborhood as opposed to any neighborhood in a city. Again, based on conceptual considerations, you could contextualize the response by asking, "Do you feel safe walking the streets *in your own neighborhood* at night alone?" Or, you could be even more specific: "Do you feel safe walking the streets *within five blocks of your home* at night alone?"

A third way to provide a context for a question is to provide an *interpretive perspective*. When the issues being asked about are sensitive, people may be unwilling to admit them to an interviewer. With spouse abuse, for example, respondents may be reluctant to admit being either a perpetrator or a victim. The scale presented in Table 6.3 begins with instructions that provide an interpretive statement that essentially says "some people do these things" and "if you have done these things, you are not alone." This provides a way of gradually moving into some issues that may be very sensitive, and eases the feeling people may have that they will be judged negatively based on their responses. It makes it more socially acceptable to admit to a behavior that could carry some stigma.

A fourth way to provide a context is to include any *definitions of terms* that may be misunderstood or any *facts* that are essential to giving accurate answers. If it were necessary to use the words *cohabitation* or *partner* (meaning one's cohabitant) in a question, for example, it might be worth preceding the question with a brief and clear statement of what you mean by (your definition of) those words.

Sensitive Questions

Questions on sensitive or potentially embarrassing topics can be a challenge to design. Sensitive topics are not limited to things such as sexual behavior or drug abuse; some people may also consider telling someone their income

TABLE 6.3 Two Subscales (Psychological Aggression and Physical Assault) of the Revised Conflict Tactics Scale (CTS2)*

Instructions:

No matter how well a couple gets along, there are times when they disagree, get annoyed with the other person, want different things from each other, or just have spats or fights because they are in a bad mood, are tired, or for some other reason. Couples also have many different ways of trying to settle their differences. This is list of things that might happen when you have differences. Please circle how many times you did each of these things in the past year, and how many times your partner did them in the past year. If you or your partner did not do one of these things in the past year, but it happened before that, circle "7."

How often did this happen?

1 = Once in the past year

2 = Twice in the past year

3 = 3–5 times in the past year

4 = 6–10 times in the past year

5 = 11–20 times in the past year

6 = More than 20 times in the past year

7 = Not in the past year, but it did happen before

0 = This has never happened

5. I insulted or swore at my partner.	1 2 3 4 5 6	7 0
25. I called my partner fat or ugly.	1 2 3 4 5 6	7 0
29. I destroyed something belonging to my partner.	1 2 3 4 5 6	7 0
35. I shouted or yelled at my partner.	1 2 3 4 5 6	7 0
49. I stomped out of the room or house or yard during a disagreement.	1 2 3 4 5 6	7 0
65. I accused my partner of being a lousy lover.	1 2 3 4 5 6	7 0
67. I did something to spite my partner.	1 2 3 4 5 6	7 0
69. I threatened to hit or throw something at my partner.	1 2 3 4 5 6	7 0
7. I threw something at my partner that could hurt.	1 2 3 4 5 6	7 0
9. I twisted my partner's arm or hair	1 2 3 4 5 6	7 0
17. I pushed or shoved my partner.	1 2 3 4 5 6	7 0
21. I used a knife or a gun on my partner.	1 2 3 4 5 6	7 0
27. I punched or hit my partner with something that could hurt.	1 2 3 4 5 6	7 0
33. I choked my partner.	1 2 3 4 5 6	7 0
37. I slammed my partner against a wall.	1 2 3 4 5 6	7 0
43. I beat up my partner.	1 2 3 4 5 6	7 0
45. I grabbed my partner.	1 2 3 4 5 6	7 0
53. I slapped my partner.	1 2 3 4 5 6	7 0
61. I burned or scalded my partner on purpose.	1 2 3 4 5 6	7 0
73. I kicked my partner.	1 2 3 4 5 6	7 0

*** Note** (1) The first eight items are the psychological aggression subscale. The numbers indicate the order in which the items would appear on the scale, but items of other subscales would be interspersed among them. Also, this table does not show the items where the respondent indicates that their partner did this to them (e.g. for Item 5, the next item would be: "My partner insulted or swore at me.") (2) Permission to use this test must be obtained from the copyright owners (Straus et al.).

Source Adapted from Murray A. Straus, Sherry L. Hamby, Sue Boney-McCoy, and David B. Sugarman, "The Revised Conflict Tactics Scale (CTS2): Development and Preliminary Psychometric Data," *Journal of Family Issues*, 17 (May 1996), 310–312.

or other aspects of their lifestyle to be private and sensitive. The problem is that the respondent may prefer to keep some things private, may be too embarrassed to respond completely, or may not know exactly how to respond, in terms of using language that will be acceptable to the interviewer.

One way of handling this problem is to provide respondents with acceptable wording for their answers in the question itself. In a question on sexual practices, for example, the response alternatives could be provided in the question itself: "We want to ask you about sexual behaviors that people sometimes engage in. People are known to engage in vaginal intercourse, oral intercourse, and anal intercourse . . ." In this way, the topic is defused to an extent because the terminology is out in the open, and the appropriate vocabulary to use for various activities is suggested. Another way to do this is to have the response alternatives listed by letters on a card; hand that card to the respondent, and ask: "Tell me the letter of the activities that you have engaged in." An illustration of this with the variable *income* is presented in Chapter 2, Table 2.2 (p. 35). This procedure is sometimes used with a variable such as *income* because people are less reluctant to tell an interviewer a letter than they are to tell the interviewer their exact income.

Response Formats

All efforts at careful wording of questions will be for naught unless the questions are presented in a manner that facilitates response. The goal is to make responding to the questions as straightforward and convenient as possible and to reduce the amount of data lost because of uninterpretable responses.

When presenting response alternatives for closed-ended questions, best results are obtained by having respondents indicate their selection by placing an X in a box corresponding to that alternative, as illustrated in question 1 of Table 6.1. Most word processing programs permit you to format a bulleted list of alternatives where the bullet can be designed in various ways, including an empty box. If a word processing program won't permit such boxes or a typewriter is being used, then the response area can be delimited with open and close brackets [ ], or parentheses ( ), with space between. These formats are preferable to open blanks and check marks (✓) because it is too easy for respondents to get sloppy and place check marks between alternatives, rendering their responses unclear and therefore useless as data. Boxes increase the likelihood that respondents will give unambiguous responses. As an alternative to X marks in boxes, it generally works well to number each response alternative and have the respondents circle the number of their choice.

When all or part of a question is closed ended, good questions also depend on providing the respondent with appropriate response alternatives that reflect the complete range of responses that might be chosen in answer to the question. In some cases, the response alternatives are appropriately either "Yes" or "No," as in question 3 in Table 6.1. Oftentimes, however, variables involve a range of opinions or attitudes. If we ask people if they favor campaign finance reform, we could offer alternatives that range from "Favor it very much" to "Favor it not at all." But does this range provide all the positions people might hold on this issue? For example, if someone feels "Strongly opposed" to campaign finance reform, does the alternative "Favor it not at all" really encapsulate that position? A more appropriate wording of a range of alternatives might be: "Favor strongly," "Favor moderately," "Unsure," "Oppose moderately," "Oppose strongly." In the first way of designing the response alternatives, all the choices include the positive word (*favor*) but that may not communicate opposition, which is a legitimate position someone could take. In Table 6.4, a number of illustrations of response formats for survey questions are presented to give you an

TABLE 6.4 A Variety of Response Formats for Survey Statements

Measuring a Student's Evaluation of a College Instructor

Overall, I rate the quality of this professor as:

□ Excellent □ Good □ Fair □ Poor

Measuring Interpersonal Conflict

How many times in the past 12 months have you insulted or swore at him or her?

□ Once □ Twice □ 3–5 times □ 6–10 times □ 11–20 times □ More than 20 times

Measuring Job Anxiety

How likely do you think it is that you will lose your job or be laid off?

□ Very likely □ Fairly likely □ Not too likely □ Not at all likely

Measuring Job Satisfaction

On the whole, how satisfied are you with the work that you do?

□ Very satisfied □ Moderately satisfied □ Moderately dissatisfied □ Very dissatisfied

Measuring Attitude Toward Military Service

For most young women, do you think military service is a good experience or not?

□ Definitely good □ Probably good □ Probably not good □ Definitely not good

Measuring Characteristics of Jobs

How important do you personally consider job security as a characteristic of a job?

Unimportant **Important**

1 2 3 4 5 6 7

Measuring Delinquency

Have you ever skipped school without a legitimate excuse?

□ Never □ Once or Twice □ Several times □ Often □ Very often

idea of some of the variations possible. (As a test of what you have learned in this section of the chapter, you might consider whether any of the statements in Table 6.4 are poorly worded and see if you can improve on their wording.)

When considering the number of response alternatives to provide, research suggests that questions or statements with more response alternatives are more reliable and valid measures than those with fewer alternatives (Alwin, 1997). Although there is considerable variation in how many alternatives should be used, five alternatives is a common number, three alternatives is probably too few, and seven or more may be better.

Real Attitudes versus Nonattitudes

When asking questions that measure people's subjective state or their attitudes, researchers often want to provide an opportunity for people to state that they don't have an attitude on an issue or that they are uncertain about how they feel. This is based on the reasonable theoretical assumption that some persons may not have an attitude about a particular issue or that they have an opinion but it is confused or uncertain. A common way of handling this is to provide a "Don't know" or "Uncertain" response alternative among those available to the respondent. If the response alternatives involve an intensity level, such as from "Strongly agree" to "Strongly disagree," then the "Uncertain" alternative is sometimes placed at the middle of the intensity range or is placed away from the other alternatives. Placing the "Uncertain" alternative in the middle implies that it is a middle position on the range of alternatives when a person's choice of "Uncertain" may not reflect that. An individual, for example, could feel strongly about an issue but may also possess a high degree of uncertainty. Setting the "Uncertain" alternative off to the side allows this expression of attitude to be clearer.

The "Don't know" and "Uncertain" response alternatives raise another problem: whether you are measuring real attitudes or nonattitudes (Converse, 1970; Gilljam & Granberg, 1993). A real attitude is, of course, one that a person actually holds and is measured by your question. A nonattitude is an attitude that is expressed but not really held by a person; it is what is called a *false positive.* It may be expressed, for example, because the design of a question does not permit the person to avoid expressing it. If there is no "Don't know" or "Uncertain" alternative, then the person is forced to express an attitude (or to not answer the question). This would suggest the importance of including such response alternatives, but there is an opposite problem: *False negatives* are attitudes that a person actually holds but are not expressed. People who do possess an attitude may choose the "Don't know" alternative because it is easier and quicker to do so.

What we do know about this problem is that both false positives and false negatives occur when assessing people's attitudes. What we know much less about is how to rectify the problem. One of the difficulties is that designing questions to reduce one problem often increases the other. Therefore, a first step is to decide which type of error is more important to avoid. For a variety of reasons, researchers are often more concerned about false positives, and these can be reduced by including a "Don't know" or "Uncertain" alternative or by asking a *filter question.* A filter question's answer determines which question the respondent goes to next. In Table 6.1, questions 2 and 3 are both filter questions. In question 2, the part of the question asking "How many items have you taken" is called a *contingency question* because whether a person answers it depends on (is contingent upon) their answer to the filter question. Note the two ways filter questions can be designed. With question 2, the person answering "Yes" is directed to their next question by the arrow and the question is clearly set off by a box; also in the box, the phrase "If Yes" is included to be sure the person realizes that this question is only for those who answered "Yes" to the previous question. With question 3, the answer "No" is followed by a statement telling the person which question they should answer next. Either format is acceptable, but again the point is to provide clear directions for the respondent.

When measuring attitudes, we sometimes use filter questions that ask persons if they have an attitude on the issue of interest as a way of reducing false positives. If the respondents state that they have an opinion, then they are presented with the question that measures their position on that issue. If false negatives are of more concern, then these options will not be included, which forces respondents to express an opinion but increases the likelihood of false positives.

Whichever of these alternatives we adopt will, of course, produce some error, and for this reason some survey researchers recommend great caution in interpreting the meaning of responses to questions. A more complex way to deal with this problem is to ask a series of questions, with the earlier ones focusing on false positives and the later ones on false negatives. A study of attitudes toward nuclear power in Sweden did this by first asking a filter question that has an "easy out" alternative (Gilljam & Granberg, 1993):

> There are various views regarding nuclear power as an energy source. What is your view? Are you generally for or against the use of nuclear power as an energy source in Sweden, or don't you have any particular opinion on this question?

A second question did not include a filter or "Don't know" alternative but did include a clear neutral point on the scale of opinion; this forces an opinion but also allows an "uncertain" type of response.

> I want to ask your position on nuclear power. Where would you place yourself on this scale?
>
> −5 −4 −3 −2 −1 0 +1 +2 +3 +4 +5

Very negative toward nuclear power	Very positive toward nuclear power

A third question was asked that offered five alternatives on how to use nuclear power (from "Shut down all plants now" to "Build more plants"), but with no neutral position. This question forces a statement of opinion with none of the easy outs. Each of these three questions is slightly different, and respondents were asked all three. This enabled the researchers, in the data-analysis phase, to assess how much response is real opinion and how much is false positives or false negatives. A final alternative to alleviate this problem is to follow up a question with an open-ended question or an in-depth interview where a respondent's attitude could be probed at some length to assess his or her real attitude.

So, special care must be taken in designing response alternatives; consideration must be given to the underlying concept being measured. Clarifying the concept can often help one design a valid set of response alternatives. In addition, pretests and tests of validity and reliability may help discover any difficulties with the alternatives. A number of additional issues relative to designing response alternatives are discussed in the remainder of this chapter while discussing index and scale construction.

Multiple-Item Indexes and Scales

Some social phenomena are too abstract or complex to be measured accurately by an individual's response to a single question or statement. When this is so, social scientists turn to multiple-item measuring devices that produce a quantitative score that is a composite of the subject's responses to a number of separate items.

Indexes versus Scales

Social scientists use two different kinds of multiple-item measuring devices: *indexes* and *scales* (DeVellis, 1991; Zeller & Carmines, 1980). However, be forewarned at the outset that some confusion surrounds the use of these two terms. Some people use them interchangeably, whereas others distinguish between them but then disagree over whether a particular measuring device is an index or a scale. In addition, some measuring devices may have properties of both. It is worth exploring these terms because they are widely used, and understanding the distinction between them will help you understand more about multiple-item measuring devices and how they are constructed. With these qualifications, let's explore what they are and examples of each.

An **index** is a composite measure in which separate indicators of a phenomenon are combined to create a single measure. In some cases,

TABLE 6.5 An example of the Bogardus Social Distance (SD) Scale

Check the "Yes" box to all of the following statements with which you agree.

I would be willing to have Norwegians as my close kin by marriage.

☐ Yes

I would be willing to have Norwegians in my club as personal friends.

☐ Yes

I would be willing to have Norwegians on my street as neighbors.

☐ Yes

I would be willing to have Norwegians working alongside me on my job.

☐ Yes

I would be willing to have Norwegians as citizens in my country.

☐ Yes

I would be willing to have Norwegians as visitors to my country.

☐ Yes

scores on each individual indicator are summed to give an overall score on the composite phenomenon. Suppose, for example, that we wanted to construct an index of involvement in delinquent activities. We could create a number of separate items that reflect particular instances of delinquent behavior, such as: Has the person been truant from school in the past year? Has the person shoplifted items from a store in the past year? Has the person vandalized any property in the past year? If each item had only two alternatives—Yes/No—we could sum up the number of Yes responses as our index of the level of delinquent activity. If there were 10 items, then a low score of 0 would represent *no delinquent activity* on our index, whereas a high score of 10 would mean that a person chose a "Yes" option to all 10 items in the index. Indexes achieve ordinal, and in some cases interval-ratio, level of measurement, as does this one. This multiple-item measure enables us to measure something more abstract (*extent of involvement in delinquency*) than the specific behaviors identified in the separate items (truancy, shoplifting, and vandalism).

A **scale** is a multiple-item measuring device in which there is a built-in intensity structure, potency, or natural levels of feeling to the items that make up the scale. A scale is made up of separate items or indicators, as is an index, but in a scale the variation in intensity among the items means that there tends to be more pattern to people's responses to the various items. Table 6.5 contains a scale that illustrates these properties. It is a version of what is called the "Bogardus Social Distance Scale," and measures people's willingness to interact with others who are different from them in terms of race, ethnicity, or culture (Converse, 1987). Like an index, a scale measures an abstract phenomenon (in this case, social distance) by combining people's responses to much more concrete and specific items (such as, "have a person as a neighbor"). In addition, if you review the items, you can see that, with some items, many people would probably agree ("willing to have Norwegians as citizens in my country"), whereas with other items, fewer people might agree. Also, an inherent order seems to exist among the items, in the sense that a "Yes" response to some items

TABLE 6.6 Rosenberg Self-Esteem Scale

	1	2	3	4
	Strongly Agree	**Agree**	**Disagree**	**Strongly Disagree**
(1) On the whole, I am satisfied with myself.	SA[4]	A[3]	D[2]	SD[1]
(2) At times, I think I am no good at all.	SA[1]	A[2]	D[3]	SD[4]
(3) I feel that I have a number of good qualities.	SA[4]	A[3]	D[2]	SD[1]
(4) I am able to do things as well as most other people.	SA[4]	A[3]	D[2]	SD[1]
(5) I feel I do not have much to be proud of.	SA[1]	A[2]	D[3]	SD[4]
(6) I certainly feel useless at times.	SA[1]	A[2]	D[3]	SD[4]
(7) I feel that I'm a person of worth, at least on an equal plane with others	SA[4]	A[3]	D[2]	SD[1]
(8) I wish I could have more respect for myself.	SA[1]	A[2]	D[3]	SD[4]
(9) All in all, I am inclined to feel that I am a failure.	SA[1]	A[2]	D[3]	SD[4]
(10) I take a positive attitude toward myself.	SA[4]	A[3]	D[2]	SD[1]

Source Morris Rosenberg, *Conceiving the Self* (rev. ed.) Malabar, FL: Krieger, Publishing (1986).

probably means that the person responded "Yes" to other items. For example, if you would admit someone as your neighbor, you would probably also admit that person to be a citizen in your country. This is the intensity structure to the items of a scale that distinguish it from an index where such an intensity structure doesn't exist. With scales, people's composite score tends to represent more of a pattern in the responses. Indexes, without the intensity structure, tend not to show such patterns; with the delinquency index, if you had answered "Yes" to shoplifting, that doesn't mean that you had also responded "Yes" to truancy or vandalism.

As we look at particular indexes and scales in this and other chapters, consider whether each has the properties of an index, a scale, or some combination of the two. Don't be misled by what others *call* a multiple-item measuring device they are using: Just because someone labels something a scale or index doesn't mean that it necessarily is. As I said, the two terms are not always used consistently.

Advantages of Multiple-Item Measures

Indexes and scales have four major advantages over single-item measures.

Improved Validity

When measuring abstract or complex variables, a multiple-item measure is generally more valid than a single-item measure. Consider the variable *self-esteem*. The Rosenberg Self-esteem Scale, developed to measure *self-esteem*, contains 10 statements (see Table 6.6). It does so because no single question or statement could possibly measure something as complex, multifaceted, and constantly changing as a person's self-esteem. What single question could be asked that might encompass all the feelings that you have about yourself? Clearly, self-esteem involves many aspects of a person's life situation—family, occupation, financial and social status, to name a few. Multiple-item scales provide more valid measures of such complex phenomena.

Improved Reliability

In general, as shown in Chapter 5, the more items contained in a measure, the more reliable it will be. This is so because the statements comprising a scale are actually just a sample of the entire universe of statements that could have been used. A single-item measure is a sample of one, and is less likely to be representative of the universe of statements than multiple items would be. Multiple-item scales are larger samples from this universe and are therefore more likely to be representative. Being more representative, they are more reliable than single-item measures.

Reduced Measurement Error

In addition to providing a more representative sampling of the items that measure a variable, multiple-item measures also reduce the impact of measurement error, especially random error. The reason for this is that any single indicator of a variable will probably measure somewhat high or low of the actual state of the variable. As more items are added to the measuring device, the likelihood that some measures will err in the opposite direction of the original items is increased. Eventually, a point is reached where sufficient items are obtained that approximately 50% measure high and 50% low, thus canceling out the random errors—one of the goals of measurement discussed in Chapter 5.

Increased Level of Measurement

Single-item measures often produce data that are nominal or, at the very best, partially ordered. The term *partially ordered data* refers to data with a few ordered categories but with many cases falling into each category, or having the same value for the variable. Although superior to nominal, these data are less desirable than fully ordered data in which every—or nearly every—case has its own rank (see Chapter 14). Multiple-item scales are capable of producing data that are closer to fully ordered and, in some cases, interval-level data. As we saw in Chapter 5, a higher level of measurement is generally preferred because it involves more precision and increased flexibility in data analysis.

When the concept to be measured is complex, multiple-item measures offer substantial advantages for the researcher—advantages that often outweigh the difficulty of their construction.

Developing Indexes and Scales

Once it is decided to use a multiple-item measuring device, an appropriate index or scale needs to be found or developed. In many cases, indexes and scales consist of questions to which individuals respond, or statements to which they indicate their level of agreement. In other cases, especially for indexes, the measuring device may consist of items of data collected elsewhere. Earlier in this chapter, guidelines were discussed for wording questions or statements in questionnaires or interviews, and those same general rules apply to the development or selection of index and scale items. In many cases, it is possible to use a complete index or scale developed by someone else if it is a valid and reliable measure of the variables under investigation. Or you might be able to use some items, or revise some items, from previously developed scales. (When using all or part of an existing measuring device, be sure to check whether you need to get permission to use it or pay copyright fees.) A major advantage of using existing indexes or scales is that their validity and reliability have usually already been established. A second advantage is that comparisons of research findings are more direct when different research projects use the same operational definitions of variables. If two projects measure a variable in different ways and reach different conclusions, it may be because they were actually measuring *different variables*. Finally, keep in mind that if you revise or use only part of an existing measuring device, it must be subjected to validity and reliability checks. A few of the many compilations of measurement scales are listed in the "For Further Reading" section of this chapter. In addition, indexes and scales are reported in the many research journals in the behavioral sciences.

If no existing index or scale will do the job, then a new scale should be developed (DeVellis, 1991; Zeller & Carmines, 1980). While there are some differences in developing indexes and scales, a certain logic is common to both. Developing multiple-item measures generally involves the following steps:

1. developing or locating many potential items, far more than will appear in the final measure;
2. eliminating items that are irrelevant, redundant, ambiguous, or for some other reason inappropriate for measuring the variable;
3. pretesting the remaining items for validity, reliability, or other measurement checks to be described shortly;
4. eliminating items that do not pass the tests of step 3; and
5. repeating steps 3 and 4 as often as necessary to reduce the index or scale to the number of items required.

Sources of Index and Scale Items

In developing or locating items to constitute an index or scale, the researcher begins with theoretical considerations: What kind of direction or guidance is provided by the theory of which the abstract concept is a part. For example, in developing the Conflict Tactics Scale (CTS), sociologist Murray Straus and his colleagues (1996) point out that "the theoretical basis of the CTS is conflict theory. . . . This theory assumes that conflict is an inevitable part of all human association, whereas violence as a tactic to deal with conflict is not" (Straus et al., 1996, p. 284). This theoretical insight led them to the conclusion that the items for the CTS had to cover not only violent tactics but also nonviolent ones. The CTS therefore includes items measuring psychological aggression, negotiation, sexual coercion, and other nonviolent modes of conflict as well as items measuring physical assault (see Table 6.3). Conflict theory also led these researchers to realize that interpersonal conflict can be a two-way street: Both partners can engage in conflictual actions. So the CTS contains items that measure the conflict tactics by both partners in a relationship. So, you can see that careful theoretical analysis gives direction to at least the *kind* of items that might be appropriate for an index or scale.

As for finding the specific items themselves, one of the most accessible sources of items for any multiple-item measure is a researcher's own imagination. Once a concept has been developed and refined, the researcher has a pretty good idea of what is to be measured. The researcher can then generate a range of statements that seem to satisfy the criteria to be discussed. At this early stage in index and scale construction, one need not be overly concerned with honing and polishing the statements to perfection because much pretesting remains before any statement is ever seen by an actual respondent.

A second source of items is people, sometimes called "judges," who are considered to be especially knowledgeable in a particular area. If one is seeking items for a delinquency scale, for instance, it would seem reasonable to discuss the issue with juvenile probation officers and others having daily contact with delinquents. Two social psychologists used this approach to find items for a scale to measure people's tendency to manipulate others for their own personal gain (Christie & Geis, 1970). They turned to the writings of Niccolò Machiavelli, a sixteenth-century adviser to the prince of Florence, Italy. In his classic book *The Prince*, Machiavelli propounded essentially a con artist's view of the world and politics: People, according to Machiavelli, are to be manipulated for one's own benefit, in a cool and unemotional fashion. In the writings of this Florentine of four centuries ago, these social psychologists found such statements as: "It is safer to be feared than to be loved," and "Humility not only is of no service but is actually harmful." They constructed a scale made up of Machiavelli's statements, somewhat revised, and asked individuals whether they agreed with each statement. The scale is now known as the "Machiavellianism Scale" and has been used widely in social science research. This illustrates a particularly creative use of "judges" in the development of multiple-item measures.

A third source of index and scale items is the persons who are the focus of the research project. Claudia Coulton (1979), for example, was interested in person-environment fit among consumers of hospital social services. In developing her scale, she obtained a large number of verbatim statements from hospital patients and then began to form these into a scale. In like manner, if one were interested in attitudes among teenagers toward unwanted pregnancies, an excellent beginning would be to discuss the topic with teenagers themselves and gather from them as many statements as possible regarding the issue. When items are garnered from individuals in this fashion, only rarely would statements be usable without editing. Many statements would ultimately be rejected, and most would have to be considerably rewritten. Such persons, however, are likely to provide a range of statements that have meaning from the perspective of the group under investigation.

Characteristics of Index and Scale Items

Once a large number of items has been found, the best have to be selected for the final measuring device. Good items have the following characteristics.

Validity

A primary concern in item selection is the validity of the statements (see Chapter 5). Each statement considered for inclusion should be assessed for face validity and content validity. For example, if we were creating a self-report delinquency scale, each statement would be assessed as to how it relates to measuring delinquent activity. Statements concerning a person's participation in various delinquent acts would be reasonable as valid measures of how delinquent that person is. On the other hand, an item relating to how well the respondent gets along with his or her siblings would probably not be a valid indicator of delinquency.

Range of Variation

Variables that are measured with multiple-item measures are normally considered to consist of a number of possible values or positions that a person could take. If we wanted to measure *attitudes toward growing old*, for example, people's positions on that variable could be extremely positive, extremely negative, or anywhere in between. In selecting items for measuring devices, we should ensure that the items cover the actual range of possible variation on the variable being measured. Failure to do so will result in a poor measuring device. When selecting items on the basis of variability, the researcher needs to exercise care to avoid defining the range either too narrowly or too broadly. Failure to include a sufficiently wide range of items will result in respondents "piling up" (clustering) at one or both ends of the scale's range. If many respondents tie with either the lowest or highest possible score, the range in the scale is inadequate. This piling up effect reduces the precision of the measurement because we are unable to differentiate among the respondents with tied scores.

Going to extremes with items to define the range is also not desirable. If we include items that are too extreme, they will apply to few, if any, respondents. In the case of a delinquency scale, for example, an item pertaining to engaging in cannibalism would be such an extreme item as to warrant exclusion. The act is so rare that it is unlikely that any juvenile has done it, and it thus contributes nothing of benefit to the scale. The goal is to select items with enough range of variation to cover the actual range of alternatives that individuals are likely to choose, without including items that are so extreme that they do not apply to anyone. In fact, one guideline is to provide a range of alternatives such that, in the long run, 50% of the respondents will choose alternatives on the positive side of the range and 50% on the negative side. However, this is only a guideline and is sometimes not necessary or possible to achieve in actual practice.

Unidimensionality

In the construction of multiple-item measures, especially scales, often the goal is to measure one specific variable. We do not want the results confounded by items on the scale that actually measure a different, although possibly related, variable. The items of a **unidimensional scale** measure only one variable. If a scale actually measures more than one variable, it is called *multidimensional.* In creating our delinquency scale, we might be tempted to include an item about school performance on the grounds that delinquents seem to perform poorly in school. Although an empirical relationship may exist between *delinquency* and *school performance,* they are separate variables and should be treated and measured as such.

Relationships between Items

To gain systematic evidence of the unidimensionality of a scale, the researcher can intercorrelate each item on the scale with every other scale item. This is often done during a pretest. If some items do not correlate with the others, it is possible that these items do not measure the same variable or they measure separate aspects of the variable that vary independently from one another. If we suspect that these items measure a different variable, they should be eliminated from the scale. If we find a few items that have nearly perfect correlations, we only need to use one of them in the scale. Two items to which individuals respond identically are merely redundant, and using both adds nothing to the measurement abilities of the scale. Occasionally, however, highly correlated items are included to detect response inconsistency or random answering. That exception notwithstanding, the final scale will be composed of statements that correlate fairly highly, but not perfectly, with one another.

A knowledge of the characteristics and sources of items provides an important and necessary foundation for the development of indexes and scales. There are many different types of indexes and scales, and the remainder of the chapter discusses them and how they are constructed.

Index Construction

One commonly used and easily understood index is the Federal Bureau of Investigation's Crime Index, which is published each year in the FBI's *Uniform Crime Reports.* The items that make up the Crime Index are the seven crimes that the FBI classifies as Part I, or more serious, crimes: homicide, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft. (Arson is also classified as a Part I offense but is not included in the Crime Index.) The Crime Index consists of the total of these seven offenses that are known to the police for every 100,000 persons (see Table 6.7).

There are several things to note about the crime index. First, each of the seven items that make it up is a part of the broader phenomenon being studied, namely, the extent of serious crimes. These items are chosen based on theoretical and conceptual considerations