Designing Questions to Be Good Measures

In: Survey Research Methods (4th ed.)

By: Floyd J. Fowler

Pub. Date: 2012
Access Date: November 10, 2017
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412958417
Online ISBN: 9781452230184
DOI: http://dx.doi.org/10.4135/9781452230184
Print pages: 86-113

Designing Questions to Be Good Measures

> **In surveys, answers are of interest not intrinsically but because of their relationship to something they are supposed to measure. Good questions are reliable (providing consistent measures in comparable situations) and valid (answers correspond to what they are intended to measure). This chapter discusses theory and practical approaches to designing questions to be reliable and valid measures.**

Designing a question for a survey instrument is designing a measure, not a conversational inquiry. In general, an answer given to a survey question is of no intrinsic interest. The answer is valuable to the extent that it can be shown to have a predictable relationship to facts or subjective states that are of interest. Good questions maximize the relationship between the answers recorded and what the researcher is trying to measure.

In one sense, survey answers are simply responses evoked in an artificial situation contrived by the researcher. The critical issue in this chapter is what an answer to a survey question tells us about some reality in which we have an interest. Let us look at a few specific kinds of answers and their meanings:

A respondent tells us that he voted for Kerry rather than Bush for president in 2004. The reality in which we are interested is which lever, if any, he pulled in the voting booth. The answer given in the survey may differ from what happened in the voting booth for any number of reasons. The respondent may have pulled the wrong lever and, therefore, not know for whom he really voted. The respondent could have forgotten for whom he voted. The respondent also could have altered his answer intentionally for some reason.

A respondent tells us how many times he went to the doctor for medical care during the past year. Is this the same number that the researcher would have come up with had he followed the respondent around for 24 hours every day during the past year? Problems of recall, of defining what constitutes a visit to a doctor, and of willingness to report accurately may affect the correspondence between the number the respondent gives and the count the researcher would have arrived at independently.

When a respondent rates her public school system as “good” rather than “fair” or “poor,” the researcher will want to interpret this answer as reflecting evaluations and perceptions of that school system. If the respondent rated only one school (rather than the whole school system), tilted the answer to please the interviewer, or understood the question differently from others, her answer may not reflect the feelings the researcher tried to measure.

Many surveys are analyzed and interpreted as if the researcher knows for certain what the answer means. Studies designed to evaluate the correspondence between respondents' answers and true values show that many respondents answer many questions very well. Even so, to assume perfect correspondence between the answers people give and some other reality is naive. When answers are good measures, it is usually the result of careful design. In the following sections, specific ways that researchers can improve the correspondence between respondents' answers and the true state of affairs are discussed.

One goal of a good measure is to increase question reliability. When two respondents are in the same situation, they should answer the question in the same way. To the extent that there is inconsistency across respondents, random error is introduced, and the measurement is less precise. The first part of this chapter deals with how to increase the reliability of questions.

There is also the issue of what a given answer means in relation to what a researcher is trying to measure: How well does the answer correspond? The later two sections of this chapter are devoted to validity, the correspondence between answers and true values and ways to improve that correspondence (Cronbach & Meehl, 1955).

Increasing the Reliability of Answers

One step toward ensuring consistent measurement is that each respondent in a sample is asked the same set of questions. Answers to these questions are recorded. The researcher would like to be able to make the assumption that differences in answers stem from differences among respondents in what they have to say rather than from differences in the stimuli to which respondents were exposed. The question's wording is obviously a central part of the stimulus.

A survey data collection is an interaction between a researcher and a respondent. In a self-administered survey, on paper or via a computer, the researcher speaks directly to the respondent through a written questionnaire or words on a computer screen. In other surveys, an interviewer reads the researcher's words to the respondent. In either case, the survey instrument is the protocol for one side of the interaction. In order to provide a consistent data collection experience for all respondents, a good question has the following properties:

- The researcher's side of the question-and-answer process is entirely scripted so that the questions as written fully prepare a respondent to answer questions.
- The question means the same thing to every respondent.
- The kinds of answers that constitute an appropriate response to the question are communicated consistently to all respondents.

Avoiding Inadequate Wording

The simplest example of inadequate question wording is when, somehow, the researcher's words do not constitute a complete question.

Incomplete Wording

Bad	*Better*
6.1 Age?	What was your age on your last birthday?
6.2 Reason last saw doctor?	What was the medical problem or reason for which you most recently went to a doctor?

Interviewers (or respondents) will have to add words or change words in order to make answerable questions from the words in the left column. If the goal is to have all respondents answering the same questions, then it is best if the researcher writes the questions fully.

Sometimes optional wording is required to fit differing respondent circumstances. That does not mean, however, that the researcher has to give up writing the questions. If the interview is computer assisted, often the computer can tailor the question wording. If a paper interview schedule is used, a common convention is to put optional wording in parentheses. These words will be used by the interviewer when they are appropriate to the situation and omitted when they are not needed.

Examples of Optional Wording

6.3 Were you (or was anyone living here with you) attacked or beaten up by a stranger during the past year?

6.4 Did (you/he/she) report the attack to the police?

6.5 How old was (EACH PERSON) on (your/his/her) last birthday?

In example 6.3, the parenthetical phrase would be omitted if the interviewer already knew that the respondent lived alone. If more than one person lived in the household, though, the interviewer would include it. The parenthetical choice offered in 6.4 may seem minor. The parentheses, however, alert the interviewer to the fact that a wording choice must be made; the proper pronoun is used, and the principle is maintained that the interviewer need read only the questions exactly as written in order to present a satisfactory stimulus.

A variation that accomplishes the same thing is illustrated in 6.5. A format such as this might be used if the same question were to be used for each person in a household. Rather than repeat the identical words endlessly, a single question is written instructing the interviewer to substitute an appropriate designation (your husband/your son/your oldest daughter).

Whether on paper or via computer, the goal is to have the interviewer ask questions that make sense and take advantage of knowledge previously gained in the interview to tailor the questions to the respondent's individual circumstances. There is another kind of optional wording that is seen occasionally in questionnaires that is not acceptable.

Example of Unacceptable Optional Wording

6.6 What do you like best about this neighborhood? (We're interested in anything, like houses, the people, the parks, or whatever.)

Presumably, this parenthetical probe was thought to be helpful to respondents who had difficulty in answering the question. From a measurement point of view, however, it undermines the principle of standardized interviewing. If interviewers use the parenthetical probe when a respondent does not readily come up with an answer, that subset of respondents will have answered a different question. Such optional probes usually are introduced when the researcher does not think the initial question is a very good one. The proper approach is to write a good question in the first place. Interviewers should not be given any options about what questions to read or how to read them except, as in the examples above, to make the questions fit the circumstances of a particular respondent in a standardized way.

The following is a different example of incomplete question wording. There are three errors embedded in the example.

Example of Poor Wording

I would like you to rate different features of your neighborhood as very good, good, fair, or poor. Please think carefully about each item as I read it.

6.7 Public schools
Parks
Public transportation
Other

The first problem with 6.7 is the order of the main stem. The response alternatives are read prior to an instruction to think carefully about the specific items. The respondent probably will forget the question. The interviewer likely will have to do some explaining or rewording before a respondent will be prepared to give an answer. Second, the words the interviewer needs to ask about the items on the list are not provided. A much better question would be the following:

> 6.7a I am going to ask you to rate different features of your neighborhood. I want you to think carefully about your answers. How would you rate (FEATURE)— would you say very good, good, fair, or poor?

The interviewer would sequentially insert each item (public schools, parks, etc.) until all four questions had been asked. This format gives the interviewer the wording needed for asking the first and all subsequent items on the list as complete questions. It also puts the elements of the question in the proper order, so that the response alternatives are read to the respondent at a point they are more likely to be remembered.

The third problem with the example is the fourth alternative, "other." What is the interviewer to say? Is he or she to make up some new question such as, "Is there anything else about your neighborhood you value?" How is the rating question to be worded? It is not uncommon to see "other" on a list of questions in a form similar to the example. Clearly, in the form presented in 6.7, the script is inadequate.

The above examples illustrate questions that could not be presented consistently to all respondents as a result of incomplete wording. Another step needed to increase consistency is to create a set of questions that flows smoothly and easily. If questions have awkward or confusing wording, if there are words that are difficult to pronounce, or if combinations of words sound awkward together, interviewers will change the words to make the questions sound better or to make them easier to read. It may be possible to train and supervise interviewers to keep such changes to a minimum. Nevertheless, it only makes sense to help interviewers by giving them questions that are as easy to read as possible.

Ensuring Consistent Meaning for All Respondents

If all respondents are asked exactly the same questions, one step has been taken to ensure that differences in answers can be attributed to differences in respondents. But there is a further consideration: The questions should all mean the same thing to all respondents. If two respondents understand the question to mean different things, their answers may be different for that reason alone.

One potential problem is using words that are not understood universally. In general population samples, it is important to remember that a range of educational experiences and cultural backgrounds will be represented. Even with well-educated respondents, using simple words that are short and understood widely is a sound approach to questionnaire design.

Undoubtedly, a much more common error than using unfamiliar words is the use of terms or concepts that can have multiple meanings. The prevalence of misunderstanding of common wording has been well documented by those who have studied the problem (e.g., Belson, 1981; Fowler, 1992; Oksenberg, Cannell, & Kalton, 1991; Tanur, 1991; Tourangeau, Rips, & Rasinski, 2000).

Poorly Defined Terms

6.8How many times in the past year have you seen or talked with a doctor about your health?

Problem. There are two ambiguous terms or concepts in this question. First, there is basis for uncertainty about what constitutes a doctor. Are only people practicing medicine with M.D. degrees included? If so, then psychiatrists are included, but psychologists, chiropractors, osteopaths, and podiatrists are not. What about physicians' assistants or nurses who work directly for doctors in doctors' offices? If a person goes to a doctor's office for an inoculation that is given by a nurse, does this count?

Second, what constitutes seeing or talking with a doctor? Do telephone consultations count? Do visits to a doctor's office when the doctor is not seen count?

Solutions. Often the best approach is to provide respondents and interviewers with the definitions they need.

We are going to ask about visits to doctors and getting medical advice from doctors. In this
6.8acase, we are interested in all professional personnel who have M.D. degrees or work
directly for an M.D. in the office, such as a nurse or medical assistant.

When the definition of what is wanted is extremely complicated and would take a very long time to define, as may be the case in this question, an additional constructive approach may be to ask supplementary questions about desired events that are particularly likely to be omitted. For example, visits to psychiatrists, visits for inoculations, and telephone consultations often are underreported and may warrant special follow-up questions. Asking specific follow-up questions to make sure such events were not left out is an easy way to reduce such errors.

Poorly Defined Terms

6.9Did you eat breakfast yesterday?

Problem. The difficulty is that the definition of breakfast varies widely. Some people consider coffee and a donut anytime before noon to be breakfast. Others do not consider that they have had breakfast unless it includes a major entree, such as bacon and eggs, and is consumed before 8 a.m. If the objective is to measure morning food consumption, the results are likely to contain considerable error stemming from differing definitions of breakfast.

Solutions. There are two approaches to the solution. On the one hand, one might choose to define breakfast:

6.9a For our purposes, let us consider breakfast to be a meal, eaten before 10:00 in the morning, that includes some protein such as eggs, meat, or milk, some grain such as toast or cereal, and some fruit or vegetable, including juice. Using that definition, did you have breakfast yesterday?

Although this often is a very good approach, in this case it is very complicated. Instead of trying to communicate a common definition to respondents, the researcher may simply ask people to report what they consumed before 10 a.m. At the coding stage, what was eaten can be evaluated consistently to see if it meets the standards for breakfast, without requiring each respondent to share the same definition.

Poorly Defined Terms

6.10Do you favor or oppose gun control legislation?

Problem. Gun control legislation can mean banning the legal sale of certain kinds of guns, asking people to register their guns, limiting the number or the kinds of guns that people may possess, or limiting which people may possess them. Answers cannot be interpreted without assumptions about what respondents think the question means. Respondents will undoubtedly interpret this question differently.

6.10a One proposal for the control of guns is that no person who ever had been convicted of a violent crime would be allowed to purchase or own a pistol, rifle, or shotgun. Would you oppose or support legislation like that?

One could argue that this is only one of a variety of proposals for gun control. That is exactly the point. If one wants to ask multiple questions about different possible strategies for gun control, one should ask separate specific questions that can be understood commonly by all respondents and interpreted by researchers. One does not solve the problem of a complex issue by leaving it to the respondents to decide what question they want to answer.

There is a potential tension between providing a complicated definition to all respondents and trying to keep questions clear and simple. This is particularly true for interviewer-administered surveys, as long definitions are particularly hard to grasp when they are delivered orally.

A potential approach is to tell interviewers to provide definitions to respondents who ask for clarification or appear to misunderstand a question. One concern about such approaches is that interviewers will not give consistent definitions if they have to improvise. However, computer-assisted interviewing makes it easy to provide interviewers with a precisely worded definition. The other, more important, concern is that only some respondents will get the needed definition. Those respondents who do not ask for clarification or appear confused will lack important information that might affect their answers.

Conrad and Schober (2000) experimented with giving interviewers freedom to provide definitions and explanations when they seemed needed. There was some evidence that accuracy improved, but the increases came at a price of more interviewer training and longer interviews. While there is need for more research on how to ask questions about complex concepts, the general approach of avoiding complex or ambiguous terms, and defining those that are used in the question wording, is the best approach for most surveys.

Avoiding Multiple Questions

Another way to make questions unreliable is to ask two questions at once.

6.11Do you want to be rich and famous?

The problem is obvious: rich and famous are not the same. A person could want to be one but not the other. Respondents, when faced with two questions, will have to decide which to answer, and that decision will be made inconsistently by different respondents.

Most multiple questions are somewhat subtler, however.

6.12 In the last 30 days, when you withdrew cash from an ATM, how often did you withdraw less than $25—always, usually, sometimes, never?

This question requires three cognitive calculations: calculate the number of visits to an ATM, the number of times less than $25 was withdrawn, and the relationship between the two numbers. While technically there is only one question, it is necessary to answer at least two prior questions in order to produce the answer. It would be better question design to use two questions.

6.12aIn the last 30 days, how many times did you withdraw cash from an ATM?
6.12b(IF ANY) On how many of those times did you withdraw less than $25?

Note two other virtues of the 6.12a and 6.12b series. First, it identifies those who did not use an

ATM at all, to whom the question does not apply. Second, by asking for numbers in both questions, it avoids having respondents do a calculation. Simplifying the demands on respondents is almost always a good idea.

6.13To what kind of place do you go for your routine medical care?

This question assumes that all respondents get routine medical care, which is not an accurate assumption. It should be asked as two questions. Probably the best approach is to ask if the respondent has gotten any routine medical care in some period—for example, the past 12 months. If so, follow with a question about the kind of place.

The “Don't Know” Option

When respondents are being asked questions about their own lives, feelings, or experiences, a “don't know” response is often a statement that they are unwilling to do the work required to give an answer. On the other hand, sometimes we ask respondents questions concerning things about which they legitimately do not know. As the subject of the questions gets further from their immediate lives, it is more plausible and reasonable that some respondents will not have adequate knowledge on which to base an answer or will not have formed an opinion or feeling. In those cases, we have another example of a question that actually is two questions at once: do you have the information needed to answer the question and, if so, what is the answer?

There are two approaches to dealing with such a possibility. One simply can ask the questions of all respondents, relying on the respondent to volunteer a “don't know” answer. Respondents differ in their willingness to volunteer that they “don't know,” however (Schuman & Presser, 1981), and interviewers are inconsistent in how they handle “don't know” responses (Fowler & Mangione, 1990; Groves, 1989). The alternative is to ask all respondents a standardized screening question about whether or not they feel familiar enough with a topic to have an opinion or feeling about it.

When a researcher is dealing with a topic about which familiarity is high, whether or not a screening question for knowledge is asked is probably not important. When a notable number of respondents will not be familiar with, or have not thought about, whatever the question is dealing with, it probably is best to ask a screening question about familiarity with the topic.

Specialized Wording for Special Subgroups

Researchers have wrestled with the fact that the vocabularies in different subgroups of the population are not the same. One could argue that standardized measurement actually would require different questions for different subgroups (Schaeffer, 1992).

Designing different forms of questionnaires for different subgroups, however, is almost never done. Rather, methodologists tend to work very hard to attempt to find wording for questions that has consistent meaning across an entire population. Even though there are situations where a question wording is more typical of the speech of one segment of a community than another (most often the better-educated segment), finding exactly comparable words for some other group of the population and then giving interviewers reliable rules for deciding when to ask which version is so difficult that it is likely to produce more unreliability than it eliminates.

The extreme challenge is how to collect comparable data from people who speak different languages. The most careful efforts translate an original version into the new language, have a different translator back translate the new version into the original language, and then try to reconcile the differences between the original and the back-translated version.

This process would be greatly improved if the designers of the original questions were concerned about ease of translation. For example, numbers translate more readily across languages than adjectives. Abstract concepts and words that are colloquial are likely to be particularly hard to translate accurately. Even when great care is taken, it is very hard to be sure people are answering comparable questions across languages. It is doubtful that adjectival rating scales are ever comparable across languages. The more concrete the questions, the better the chances for comparability of results across languages or cultures. Marin and Marin (1991) present a good analysis of the challenges of collecting comparable data from English and Spanish-speaking people. Harkness, Van de Vijver, and Mohler (2003) provide a comprehensive look at the challenges of collecting comparable data across cultures.

Standardized Expectations for Type of Response

As stated, it is important to give interviewers a good script so that they can read the questions exactly as worded, and it is important to design questions that mean the same thing to all respondents. The other key component of a good question is that respondents should have the same perception of what constitutes an adequate answer for the question.

The simplest way to give respondents the same perceptions of what constitutes an adequate answer is to provide them with a list of acceptable answers. Such questions are called closed questions. The respondent has to choose one, or sometimes more than one, of a set of alternatives provided by the researcher.

6.14 What was the main reason you went to the doctor—for a new health problem, for a follow-up for a previous health problem, for a routine checkup, or for some other reason?

Closed questions are not suitable in all instances. The range of possible answers may be more extensive than it is reasonable to provide. The researcher may not feel that all reasonable answers can be anticipated. For such reasons, the researcher may prefer not to provide a list of alternatives to the respondent. In that case, the question must communicate the kind of response wanted as well as possible.

6.15When did you have the measles?

Problem. The question does not specify the terms in which the respondent is to answer. Consider the following possible answers: “Five years ago;” “While I was in the army;” “When I was pregnant with our first child;” “When I was 32;” “In 1987.” All of these answers could be given by the same person, and all are appropriate answers to the question as posed. They are not all acceptable in the same survey, however, because descriptive statistics require comparable answers. An interviewer cannot use the words in example 6.14 and consistently obtain comparable data, because each respondent must guess what kind of answer is wanted.

Solution. A new question must be created that explains to the respondent what kind of answer is wanted.

6.15aHow old were you when you had the measles?

Obviously, 6.15a is the way the question should have been worded by the researcher for all respondents.

6.16Why did you vote for candidate A?

Problems. Almost all “why” questions pose problems. The reason is that one's sense of causality or frame of reference can influence answers. In the particular instance above, the respondent may choose to talk about the strengths of candidate A, the weaknesses of candidate B, or the reasons he or she used certain criteria (My mother was a lifelong Republican). Hence respondents who see things exactly the same way may answer differently.

Solution. Specify the focus of the answer:

6.16aWhat characteristics of candidate A led you to vote for (him/her) over candidate B?

Such a question explains to respondents that the researcher wants them to talk about candidate A, the person for whom they voted. If all respondents answer with that same frame of reference, the researcher then will be able to compare responses from different respondents in a direct fashion.

6.17What are some of the things about this neighborhood that you like best?

Problems. In response to a question like this, some people will only make one or two points, whereas others will make many. It is possible that such differences reflect important differences in respondent perceptions or feelings. Research has shown pretty clearly, however, that education is related highly to the number of answers people give to such questions. Interviewers also affect the number of answers.

Solution. Specify the number of points to be made:

6.17a What is the feature of this neighborhood that you would single out as the one you like most?

6.17b Tell me the three things about this neighborhood that you like most about living here.

Although this may not be a satisfactory solution for all questions, for many such questions it is an effective way of reducing unwanted variation in answers across respondents.

The basic point is that answers can vary because respondents have a different understanding of the kind of responses that are appropriate. Better specification of the properties of the answer desired can remove a needless source of unreliability in the measurement process.

Types of Measures/Types of Questions

The above procedures are designed to maximize reliability, the extent to which people in comparable situations will answer questions in similar ways. One can measure with perfect reliability, though, and still not be measuring what one wants to measure. The extent to which the answer given is a true measure and means what the researcher wants or expects it to mean is called *validity.* In this section, aspects of the design of questions are discussed, in addition to steps to maximize the reliability of questions, that can increase the validity of survey measures.

For this discussion, it is necessary to distinguish between questions designed to measure facts or objectively measurable events and questions designed to measure subjective states such as attitudes, opinions, and feelings. Even though there are questions that fall in a murky area on the border between these two categories, the idea of validity is somewhat different for objective and subjective measures.

If it is possible to check the accuracy of an answer by some independent observation, then the measure of validity becomes the similarity of the survey report to the value of some "true" measure. In theory, one could obtain an independent, accurate count of the number of times that an individual used an ATM during a year. Although in practice it may be very difficult to obtain such an independent measure (e.g., getting access to the relevant records could be impossible), the understanding of validity can be consistent for objective situations.

In contrast, when people are asked about subjective states, feelings, attitudes, and opinions, there is no objective way of validating the answers. Only the respondent has access to his or her feelings and opinions. Thus the validity of reports of subjective states can be assessed only by their correlations with other answers that a person gives or with other facts about the respondent's life that one thinks should be related to what is being measured. For such measures, there is no truly independent direct measure possible; the meaning of answers must be inferred from patterns of association.

Levels of Measurement

There are four different ways in which measurement is carried out in social sciences. This produces four different kinds of tasks for respondents and four different kinds of data for analysis:

> *Nominal*—people or events are sorted into unordered categories (Are you male or female?)
>
> *Ordinal*—people or events are ordered or placed in ordered categories along a single dimension (How would you rate your health—very good, good, fair, or poor?)
>
> *Interval data*—numbers are attached that provide meaningful information about the distance between ordered stimuli or classes (in fact, interval data are very rare; Fahrenheit temperature is one of the few common examples)
>
> *Ratio data*—numbers are assigned such that ratios between values are meaningful, as well as the intervals between them. Common examples are counts or measurements by an objective, physical scale such as distance, weight, or pressure (How old were you on your last birthday?)

Most often in surveys, when one is collecting factual data, respondents are asked to fit themselves or their experiences into a category, creating nominal data, or they are asked for a number, most often ratio data. “Are you employed?;” “Are you married?;” and “Do you have arthritis?” are examples of questions that provide nominal data. “How many times have you seen a doctor?;” “How much do you weigh?;” and “What is the hourly rate you are paid?” are examples of questions that ask respondents to provide real numbers for ratio data.

When gathering factual data, respondents may be asked for ordinal answers. For example, they may be asked to report their incomes in relatively large categories or to describe their behavior in nonnumerical terms (e.g., usually, sometimes, seldom, or never). When respondents are asked to report factual events in ordinal terms, it is because great precision is not required by the researcher or because the task of reporting an exact number is considered too difficult. There usually is a real numerical basis, however, underlying an ordinal answer to a factual question.

The situation is somewhat different with respect to reports of subjective data. Although there have been efforts over the years, first in the work of psychophysical psychologists (e.g., Thurstone & Chave, 1929), to have people assign numbers to subjective states that met the assumptions of interval and ratio data, for the most part respondents are asked to provide nominal and ordinal data about subjective states. The nominal question is, “Into which category do your feelings, opinions, or perceptions fall?” The ordinal question is “Where along this continuum do your feelings, opinions, or perceptions fall?”

When designing a survey instrument, a basic task of the researcher is to decide what kind of measurement is desired. When that decision is made, there are some clear implications for the form in which the question will be asked.

Types of Questions

Survey questions can be classified roughly into two groups: those for which a list of acceptable responses is provided to the respondent (closed questions) and those for which the acceptable responses are not provided exactly to the respondent (open questions).

When the goal is to put people in unordered categories (nominal data), the researcher has a choice about whether to ask an open or closed question. Virtually identical questions can be designed in either form.

Examples of Open and Closed Questions

6.18	What health conditions do you have? (open)	
6.18a	Which of the following conditions do you currently have? (READ LIST) (closed)	
	What do you consider to be	Here is a list of problems that many people in the

6.19	the most important problem facing our country today? (open)	6.19a	country are concerned about. Which do you consider to be the most important problem facing our country today? (closed)

There are advantages to open questions. They permit the researcher to obtain answers that were unanticipated. They also may describe more closely the real views of the respondents. Third, and this is not a trivial point, respondents like the opportunity to answer some questions in their own words. To answer only by choosing a provided response and never to have an opportunity to say what is on one's mind can be a frustrating experience. Finally, open questions are appropriate when the list of possible answers is longer than is feasible to present to respondents.

Despite all this, however, closed questions are usually a more satisfactory way of creating data. There are four reasons for this:

The respondent can perform more reliably the task of answering the question when response alternatives are given.

The researcher can perform more reliably the task of interpreting the meaning of answers when the alternatives are given to the respondent (Schuman & Presser, 1981).

When a completely open question is asked, many people give relatively rare answers that are not analytically useful. Providing respondents with a constrained number of answer options increases the likelihood that there will be enough people giving any particular answer to be analytically interesting.

Since most data collection now is computer assisted, it is much easier for interviewers or respondents to record answers by checking a provided answer than to key in narrative answers.

Finally, if the researcher wants ordinal data, the categories must be provided to the respondent. One cannot order responses reliably along a single continuum unless a set of permissible ordered answers is specified in the question. Further discussion about the task that is given to respondents when they are asked to perform an ordinal task is appropriate, because it is probably the most prevalent kind of measurement in survey research.

Figure 6.1 Subjective Continuum Scales

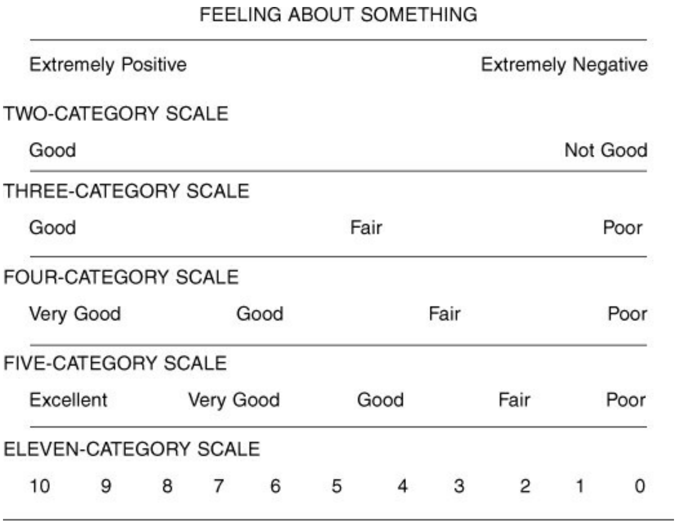

Figure 6.1 shows a continuum. (This case concerns having respondents make a rating of some sort, but the general approach applies to all ordinal questions.) There is a dimension assumed by the researcher that goes from the most negative feelings possible to the most positive feelings possible. The way survey researchers get respondents into ordered categories is to put designations or labels on such a continuum. Respondents then are asked to consider the labels, consider their own feelings or opinions, and place themselves in the proper category.

There are two points worth making about the kinds of data that result from such questions. First, respondents will differ in their understanding of what the labels or categories mean. The only assumption that is necessary in order to make meaningful analyses, however, is that, on the average, the people who rate their feelings as “good” feel more positively than those who rate their feelings as “fair.” To the extent that people differ some in their understanding of and criteria for “good” and “fair,” there is unreliability in the measurement, but the measurement still may have meaning (i.e., correlate with the underlying feeling state that the researcher wants to measure).

Second, an ordinal scale measurement like this is relative. The distribution of people choosing a particular label or category depends on the particular scale that is presented.

Consider the rating scale in Figure 6.1 again and consider two approaches to creating ordinal scales. In one case, the researcher used a 3-point scale: good, fair, or poor. In the second case, the researcher used five descriptive options: excellent, very good, good, fair, and poor. When one compares the two scales, one can see that adding “excellent” and “very good” in all probability does not simply break up the “good” category into three pieces. Rather, it changes the whole sense of the scale. People respond to the ordinal position of categories as well as to the descriptors.

A recent experiment makes the point (Wilson, Alman, Whitaker, & Callegro, 2004). Respondents were asked to use two 5-point scales to rate their health—one identical to the 5-point scale in Figure 6.1 and the other using “very good, good, moderate, bad, and very bad.” Respondents then were asked to use a scale from 1 to 10 to provide a numerical equivalent for each verbal category in the two scales.

As one would expect, the numbers given to “very good” were higher when it was the first answer (9.8 vs. 7.8) and “good” received a numerical score of 7.3 when it was the second category, but only 5.4 when it was third.

Such scales are meaningful if used as they are supposed to be used: to order people. By itself, however, a statement that some percentage of the population feels something is “good or better” is not appropriate, because it implies that the population is being described in some absolute sense. In fact, the percentage would change if the question were different. Only comparative statements (or statements about relationships) are justifiable when one is using ordinal measures:

- comparing answers to the same question across groups (e.g., 20% more of those in group A than in group B rated the candidate as “good or better”); or
- comparing answers from comparable samples over time (e.g., 10% more rated the candidate “good” or better in January than did so in November).

The same general comments apply to data obtained by having respondents order items (e.g., Consider the schools, police services, and trash collection. Which is the most important city service to you?). The percentage giving any item top ranking, or the average ranking of an item, is completely dependent on the particular list provided. Comparisons between distributions when the alternatives have been changed at all are not meaningful.

Agree-Disagree Items: A Special Case

Agree-disagree items are very prevalent in survey research and therefore deserve special attention. The task that respondents are given in such items is different from that of placing themselves in an ordered category. The usual approach is to read a statement to respondents and to ask them if they agree or disagree with that statement. The statement is located somewhere on a continuum such as that portrayed in Figure 6.1. Respondents' locations on that continuum are calculated by figuring out whether they say their feelings are very close to that statement (by agreeing) or are very far from where that statement is located (by disagreeing).

When one compares questions posed in the agree-disagree format with questions in the straightforward rating format, there are numerous disadvantages to the former. Compare the following:

6.20 My health is poor. Do you strongly agree, agree, disagree, or strongly disagree?
6.20aHow would you rate your health—excellent, very good, good, fair, or poor?

The disadvantages to the first statement are as follows:

- The rating scale sorts respondents into five categories; the agree-disagree question is almost always analyzed by putting respondents into two groups (agrees or disagrees). Hence more information is gained from the rating.
- Agree-disagree questions, in order to be interpretable, can only be asked about extremes of a continuum. If the statement was, “My health is fair,” a person could disagree either because it was “good” or because it was “poor.” This feature limits the ability to order people in the middle of a continuum.
- Respondents often find it confusing that the way to say their health is good is to disagree that their health is poor.
- Studies show that some respondents are particularly likely to agree (or acquiesce) when questions are put in this form; that is, there are people who would agree both that their health is “poor” and that it is “not poor” if question 6.20 was stated in the negative (Dillman & Tarnai, 1991; Krosnick, Judd, & Wittenbrink, 2007; Schuman & Presser, 1981).

Because of these complexities, it is routinely found that the direct rating task has more validity than the comparable agree-disagree question. For unidimensional scaling tasks, it is hard to justify using 6.20 rather than 6.20a. A very common usage of the format, however, is to obtain responses to complex statements such as the following:

6.21 With economic conditions the way they are these days, it really isn't fair to have more than two children.

This question is asking about at least three things at once: the perceived state of the economy, views on the appropriate maximum number of children, and views about the relationship between the economy and family size.

If a person does not happen to think that economic conditions are bad (which the question imposes as an assumption) and/or that economic conditions of whatever kind have any implications for family size, but if that person happens to think two children is a good target for a family, it is not easy to answer the question. Moreover, whether a person agrees or disagrees, it is hard to know what the respondent agreed or disagreed with.

The agree-disagree format appears to be a rather simple way to construct questions. In fact, to use this form to provide reliable, useful measures is not easy and requires a great deal of care and attention. Usually, researchers will have more reliable, valid, and interpretable data if they avoid the agree-disagree question form.

Increasing the Validity of Factual Reporting

When a researcher asks a factual question of a respondent, the goal is to have the respondent report with perfect accuracy; that is, give the same answer that the researcher would have given if the researcher had access to the information needed to answer the question. There is a rich methodological literature on the reporting of factual material. Reporting has been compared against records in a variety of areas, in particular the reporting of economic and health events (see Cannell, Marquis, & Laurent, 1977, for a good summary, as well as Edwards et al., 1994; Edwards, Winn, & Collins, 1996; Tourangeau, Rips, & Rasinski, 2000).

Respondents answer many questions accurately. For example, more than 90% of overnight hospital stays within 6 months of an interview are reported (Cannell, Marquis, & Laurent, 1977). How well people report, however, depends on both what they are being asked and how it is asked. There are four basic reasons why respondents report events with less than perfect accuracy:

They do not understand the question.
They do not know the answer.
They cannot recall it, although they do know it.
They do not want to report the answer in the interview context.

There are several steps that the researcher can take to combat each of these potential problems. These steps are reviewed below.

Understanding the Question

If respondents do not all have the same understanding of what the questions ask for, error is certain to result. As discussed earlier, when researchers are trying to count events that have complex definitions, such as burglaries or physician services, they have two options: (a) Provide definitions to all respondents; or (b) have respondents provide the information needed to classify their experiences into detailed, complex categories, and then have coders categorize answers.

Fowler (1992) has shown that people do answer questions that include ambiguous terms, producing quite distorted data. Researchers cannot assume that respondents will ask for clarification if they are not sure what a question means. To maximize the validity of factual survey data, an essential first step is to write questions that will be consistently understood by all respondents.

Lack of Knowledge

Lack of knowledge as a source of error is of two main types: (a) The chosen respondent does not know the answer to the question, but someone in the selected household does; or (b) no one in the selected household knows the answer. The solution in the first situation lies in choosing the right respondent, not question design. Most often, the problem is that one household respondent is asked to report information about other household members or the household as a whole. Solutions include the following:

- Identify and interview the household member who is best informed.
- Use data collection procedures that permit the respondent to consult with other household members.
- Eliminate proxy respondents; only ask respondents to provide information about themselves.

Sometimes a complex data collection strategy is called for. For example, the National Crime Victimization Survey conducted by the Bureau of the Census obtains reports of household crimes from a single household informant, but in addition asks each household adult directly about personal crimes such as robbery. If the basic interview is to be carried out in person, costs for interviews with other members of the household can be reduced if self-administered forms are left to be filled out by absent household members, or if secondary interviews are done by telephone. A variation is to ask the main respondent to report the desired information as fully as possible for all household members, then mail the respondent a summary for verification, permitting consultation with other family members.

When respondents are asked questions about themselves that they cannot answer, it is a question design problem. In theory, one could differentiate between information the respondent cannot recall and information the respondent never had at all. In either case, the problem for the researcher is to design questions that almost everyone can answer. Among the options available are the following:

- Change the question to ask for information that is less detailed or easier to recall.

- Help the respondent estimate the answer.
- Change or drop the objective.

It is not uncommon for questions to ask for answers in more detail than the research objectives require.

> The question asks respondents for the name of all the medications they take (a very hard question) when the objective is to find out who is taking medicine for high blood pressure (a much easier question).
>
> The question asks for income in an open-ended (and implicitly very detailed way) when getting an estimate of income in broad categories would satisfy the research objectives.

Recall follows some obvious principles: Small events that have less impact are more likely to be forgotten than more significant events; recent events are reported better than events that occurred in the more distant past (Cannell, Marquis, & Laurent, 1977). Sometimes it may be worthwhile to change question objectives to improve reporting by asking about events that are easier to recall. For example, although it may be desirable to have respondents report all the crimes that happened in the last year, there will be less reporting error if they are asked to report for only 6 months.

A comparatively new set of question design strategies has resulted from the growing involvement of cognitive psychologists in survey methods (Jabine, Straf, Tanur, & Tourangeau, 1984; Sirken, Herrmann, Schechter, Schwarz, Tanur, & Tourangeau, 1999; Willis, 2005). Various strategies are being tried to help respondents recall events (e.g., by suggesting possible associations) or place events in time (e.g., by having respondents recall something that happened about a year before). Event calendars help respondents place events in time and recall events by putting them in context (Belli, Lee, Stafford, & Chou, 2004).

For many survey tasks, studies have shown that respondents do not actually use recall to answer some questions; they estimate the answers (e.g., Burton & Blair, 1991). For example, if respondents are asked for the number of times they visited a grocery store to buy food in some period, they usually estimate based on their usual patterns rather than try to remember the individual events. This observation leads researchers to design strategies for helping respondents make better estimates.

Finally, it is important to recognize that there are some things that researchers would like to have people report that they cannot. For example, people do not know the cost of their medical care that is paid by insurance. If one truly wants to obtain medical costs, it is necessary to supplement what respondents may be able to report (for example, their out-of-pocket expenditures) with data collected directly from providers or insurers.

Social Desirability

There are certain facts or events that respondents would rather not report accurately in an interview. Health conditions that have some degree of social undesirability, such as mental illness and venereal disease, are underreported significantly more than other conditions. Hospitalizations associated with conditions that are particularly threatening, either because of the possible stigmas that may be attached to them or because of their life-threatening nature, are reported at a lower rate than average (Cannell, Marquis, & Laurent, 1977). Aggregate estimates of alcohol consumption strongly suggest underreporting, although the reporting problems may be a combination of recall difficulties and respondents' concerns about social norms regarding drinking. Arrest and bankruptcy are other events that have been found to be underreported consistently but seem unlikely to have been forgotten (Locander, Sudman, & Bradburn, 1976).

There are probably limits to what people will report in a standard interview setting. If a researcher realistically expects someone to admit something that is very embarrassing or illegal, extraordinary efforts are needed to convince respondents that the risks are minimal and that the reasons for taking any risk are substantial. The following are some of the steps that a researcher might consider when particularly sensitive questions are being asked (also see Catania, Gibson, Chitwood, & Coates, 1990; Sudman & Bradburn, 1982).

1. *Minimize a sense of judgment; maximize the importance of accuracy.* Careful attention to the introduction and vocabulary that might imply the researcher would value certain answers negatively is important.

Researchers always have to be aware of the fact that respondents are having a conversation with the researcher. The questions and the behavior of the interviewer, if there is one, constitute all the information the respondent has about the kind of interpretation the researcher will give to the answers. Therefore, the researcher needs to be very careful about the cues respondents are receiving about the context in which their answers will be interpreted.

2. *Use self-administered data collection procedures.* Although the data are not conclusive, there is some evidence that telephone interviews are more subject to social desirability bias than personal interviews (Aquilino, 1994; de Leeuw & van der Zouwen, 1988; Fowler et al., 1998; Henson et al., 1977; Mangione et al., 1982). The evidence is much clearer that having

respondents answer questions in a self-administered form, on paper or directly into a computer, rather than having an interviewer ask the questions will produce less social desirability bias for some items (e.g., Aquilino, 1994; Aquilino & Losciuto, 1990; Dillman & Tarnai, 1991; Fowler et al., 1998; Hochstim, 1967). For surveys dealing with sensitive topics, a mail survey or group administration should be considered. A personal interview survey also can include some self-administered questions: A respondent simply is given a set of questions to answer in a booklet. If the survey is computer-assisted, the respondents can enter their answers directly into a computer with much the same effect. For example, such an approach has been shown to significantly increase reports of recent illegal drug use (Penne, Lessler, Beiler, & Caspar 1998; Tourangeau & Smith, 1998). Finally, Turner and colleagues (1998) and Villarroel and colleagues (2006) have shown that telephone surveys obtain much higher estimates of socially sensitive activities related to sex and drugs when answers are entered directly into a computer using the touch-tone feature on the telephone than when an interviewer asks the questions.

3. *Assure confidentiality and anonymity.* Almost all surveys promise respondents that answers will be treated confidentially and that no one outside the research staff will ever be able to associate individual respondents with their answers. Respondents usually are assured of such facts by interviewers in their introductions and in advance letters, if there are any; these may be reinforced by signed commitments from the researchers. Self-administered forms that have no identifiers provide a way to ensure that answers are anonymous— not just confidential. Finally, for surveys on particularly sensitive or personal subjects, there are some elaborate survey strategies, such as random response techniques, that respondents cannot be linked to their answers (These are described by Fox & Tracy, 1986, and by Fowler, 1995.)

Again it is important to emphasize that the limit of survey research is what people are willing to tell researchers under the conditions of data collection designed by the researcher. There are some questions that probably cannot be asked of probability samples without extraordinary efforts. Some of the procedures discussed in this section, however, such as trying to create a neutral context for answers and emphasizing the importance of accuracy and the neutrality of the data collection process, are probably worthwhile procedures for the most innocuous of questions. Any question, no matter how innocent it may seem, may have an answer that is embarrassing to somebody in the sample. It is best to design all phases of a survey instrument with a sensitivity to reducing the effects of social desirability and embarrassment for any answers people may give.

Increasing the Validity of Answers Describing Subjective States

As discussed above, the validity of subjective questions has a different meaning from that of objective questions. There is no external criterion; one can estimate the validity of a subjective measure only by the extent to which answers are associated in expected ways with the answers to other questions, or other characteristics of the individual to which it should be related (see Turner & Martin, 1984, for an extensive discussion of issues affecting the validity of subjective measures).

There basically are only three steps to the improvement of validity of subjective measures:

> Make the questions as reliable as possible. Review the sections on the reliability of questions, dealing with ambiguity of wording, standardized presentation, and vagueness in response form, and do everything possible to get questions that will mean the same thing to all respondents. To the extent that subjective measures are unreliable, their validity will be reduced. A special issue is the reliability of ordinal scales, which are dominant among measures of subjective states. The response alternatives offered must be unidimensional (i.e., deal with only one issue) and monotonic (presented in order, without inversion).

Problematic Scales

6.22 How would you rate your job—very rewarding, rewarding but stressful, not very rewarding but not stressful, or not rewarding at all?

6.23 How would you rate your job—very rewarding, somewhat rewarding, rewarding, or not rewarding at all?

Question 6.22 has two scaled properties, rewardingness and stress, that need not be related. Not all the alternatives are played out. Question 6.22 should be made into two questions if rewardingness and stress of jobs are both to be measured. In 6.23, some would see “rewarding” as more positive than “somewhat rewarding” and be confused about how the categories were ordered. Both of these problems are common and should be avoided.

> When putting people into ordered classes along a continuum, it probably is better to have more categories than fewer. There is a limit, however, to the precision of discrimination that respondents can exercise in giving ordered ratings. When the number of categories exceeds the respondents' ability to discriminate their feelings, numerous categories simply produce unreliable noise. Also numerous categories may make questions harder to administer, particularly on the telephone. However, to the extent that real variation among respondents is being measured, more categories will increase validity.
>
> Ask multiple questions, with different question forms, that measure the same subjective state; combine the answers into a scale. The answers to all questions potentially are influenced both by the subjective state to be measured and by specific features of the respondent or of

the questions. Some respondents avoid extreme categories; some tend to agree more than disagree. Multiple questions help even out response idiosyncrasies and improve the validity of the measurement process (Cronbach, 1951; DeVellis, 2003).

The most important point to remember about the meaning of subjective measures is their relativity. Distributions can be compared only when the stimulus situation is the same. Small changes in wording, changing the number of alternatives offered, and even changing the position of a question in a questionnaire can make a major difference in how people answer (see Schuman & Presser, 1981; Sudman & Bradburn, 1982; and Turner & Martin, 1984, for numerous examples of factors that affect response distributions). The distribution of answers to a subjective question cannot be interpreted directly; it only has meaning when differences between samples exposed to the same questions are compared or when patterns of association among answers are studied.

Question Design and Error

A defining property of social surveys is that answers to questions are used as measures. The extent to which those answers are good measures is obviously a critical dimension of the quality of survey estimates. Questions can be poor measures because they are unreliable (producing erratic results) or because they are biased, producing estimates that consistently err in one direction from the true value (as when drunk driving arrests are underreported).

We know quite a bit about how to make questions reliable. The principles outlined in this chapter to increase reliability are probably sound. Although other points might be added to the list, creating unambiguous questions that provide consistent measures across respondents is always a constructive step for good measurement.

The validity issue is more complex. In a sense, each variable to be measured requires research to identify the best set of questions to measure it and to produce estimates of how valid the resulting measure is. Many of the suggestions to improve reporting in this chapter emerged from a 20-year program to evaluate and improve the measurement of health-related variables (Cannell, Marquis, & Laurent, 1977; Cannell, Oksenberg, & Converse, 1977). There are many areas in which a great deal more work on validation is needed.

Reducing measurement error through better question design is one of the least costly ways to improve survey estimates. For any survey, it is important to attend to careful question design and pretesting (which are discussed in Chapter 7) and to make use of the existing research literature about how to measure what is to be measured. Also, continuing to build a literature in which the validity of measures has been evaluated and reported is much needed. Robinson, Shaver, and Wrightsman (1997) and McDowell (2003) have compiled data on the validity of many commonly used multi-item measures that document how measures have been validated, as well as how much work remains to be done.

Exercises

Use the criteria discussed in this chapter to evaluate the following questions as reliable, interpretable, and analytically useful measures; write better questions if you can.

a. To measure income: How much do you make?

b. To measure health: How healthy are you?

c. To measure satisfaction with life: How would you rate your life—very good, better than average, mixed, could be better, or very bad?

d. To measure opinion about abortion laws: Tell me whether you agree or disagree with the following statement: Abortion is morally very questionable; abortions should be illegal, except in emergencies.

Write a series of questions to measure position for or against universal health insurance.

Write a series of questions to measure degree of political involvement.

Write a hypothesis about a possible relationship between two variables (e.g., good health is associated with receiving good quality health care; or good quality housing is related to having a high income). Then, under each part of the hypothesis, write the information you would need in order to assign a value to a person for each of the two variables. Then draft a question (or set of questions) for each part, the answers to which would provide the information you need. Indicate whether your questions ask for factual or subjective information and whether the resulting data will have nominal, ordinal, interval, or ratio properties.

Further Readings

http://dx.doi.org/10.4135/9781452230184.n6