Questionnaire Design and Development

Next question: I believe that life is a constant striving for balance, requiring frequent tradeoffs between morality and necessity, within a cyclic pattern of joy and sadness, forging a trail of bittersweet memories until one slips, inevitably, into the jaws of death. Agree or disagree?


This chapter is intended to introduce you to methods for developing questionnaires for use in surveys. The term *survey research* describes studies designed to collect observations about social phenomena in a systematic manner through interviews and questionnaires administered to samples of individuals. Survey results are widely used by academic institutions, businesses, and government agencies; they have come to play an important role in journalism and legal proceedings as well. Estimates derived from “surveys of surveys” indicate that at least 100 million survey interviews were conducted between 1971 and 1976 in the United States, and more than 28 million interviews were performed by telephone in 1980 (Turner and Martin 1984, p. 30). Clearly, few among us have not been exposed to the ubiquitous questionnaire.

Like other crafts with low entry barriers, questionnaire design is practiced by many but mastered by few. This discussion is directed to the first-time questionnaire designer seeking to benefit from the experience of professional practitioners and social scientists. It will cover the major concerns that must be addressed in formulating a questionnaire as well as some of the design options available. A large body of literature exists on the complex subject of questionnaire design and this note can only hope to skim the guidance, ideas, and examples available there. Accordingly, at numerous points in the ensuing paragraphs, references to additional source materials will be cited.

This note is organized as follows. We begin by outlining a multistep process for developing a questionnaire. The next section discusses the types and sources of errors in survey data. Following that, we examine the problems of identifying the information to be sought and specifying the role and tasks of respondents. The final section addresses some topics related to question wording and questionnaire organization. The note concludes with a brief summary of the main points.
A Multistage Development Process

This section will describe a step-by-step process for developing a questionnaire that is similar to those proposed by Sheatsley (1983) and Sudman and Bradburn (1982). It is assumed that by this time you have carefully defined the purpose of your study and have made a tentative decision about the data collection method you will employ. However, in the course of designing the questionnaire, you may find it necessary to refine and modify your goals and approaches as new issues arise. The process of developing a questionnaire is very much an iterative one and indeed, much of the value to be realized from pursuing development in an orderly, sequential manner derives from expanding one’s consciousness about complexities and paying close attention to details that bear on the quality of the finished instrument. The following outline shows the steps involved in developing the questionnaire. These steps are explained in more detail beginning on page 8.

1. Decide What Information Should be Sought
   a) Determine what data you need to address your research purposes and questions.
   b) Translate data requirements into respondent/informant tasks.

2. Draft Questions
   a) Consult sources for design guidelines.
   b) Choose question type.
   c) Write question and check for wording problems.
   d) Design response format.
   e) Determine requirements for classification information, e.g., respondent’s demographic characteristics, and formulate questions for same.

3. Design Questionnaire
   a) Draft introduction and instructions.
   b) Order questions.
   c) Format questionnaire and code for processing.

4. Review First Draft Critically and Revise
   a) Identify omissions and excesses.
   b) Administer to a naive subject, solicit feedback, and estimate completion time required.
   c) Apply question critique checklist.

5. Pilot Test, Repeat Step #4, and Revise

Survey Error: Problems and Opportunities

Measurement of any quantity is, of course, subject to error, and estimates obtained from surveys are no exception. The quality of survey results reflects the way in which the survey was designed and executed and how the findings were analyzed. It is important to recognize the possible sources of error in survey results. The figure shown below outlines the major types of errors encountered in surveys.
The first distinction to be made is between sampling and nonsampling errors (Cochran 1977). Sampling errors arise because observations are obtained from a sample of some population rather than from a complete enumeration of that population. If the measurement process were to be repeated over and over again, each time drawing a different sample of the same size, the results obtained would exhibit some random variation from sample to sample. Thus, the news media announce that “a nationwide poll of 1,000 adults found that 63% favor passage of the Equal Rights Amendment, with a margin of error of plus or minus 3 percentage points.” The “margin of error” noted in such reports, typically 3 percentage points, ordinarily refers to estimates of sampling error.1

Nonsampling Error

Ignoring nonsampling error results in understatements of total survey error. Converse and Traugott (1986) reviewed a body of data derived from comparisons of results from similar or matched surveys of the same public policy issues. Analysis of the comparisons indicated that the discrepancies observed exceeded those expected to arise from random sampling errors—leading them to suggest that “the conventional 3 percentage point warning is rarely a conservative one” (p. 1095).

Nonsampling errors may be divided into: (a) nonresponse errors, which occur when responses from certain members of the original sample are not obtained, and (b) measurement errors, sometimes called “response effects” (Sudman and Bradburn 1974, pp. 2–4), which can be traced to the instruments used to gather observations (e.g., interviewers or questionnaires) and/or to the participants in the study.

Findings from a study conducted for AT&T illustrate these distinctions (Assael and Keon 1982). A questionnaire mailed to a sample of small businesses (defined as firms having three phone lines or less) included the following item: “What was the company’s average telephone bill over the last three months?” Of the original sample of firms to whom the questionnaire was mailed, 58% returned the questionnaire and responded to the above item. The mean bill reported was $136/firm/month. Using billing records, AT&T was able to determine the actual telephone bills for both the total sample to whom the questionnaire had been sent originally as well as the 58% who responded. These “true” (mean) values were found to be $113 for the total original sample and $127 for the sample of responding firms.

1 For a simple random sample of size n, the standard error (S.E.) of a proportion, p, is given by: S.E. = (p (1-p)/n)^1/2. Thus for p = .5 and n = 1,000, we obtain S.E. = 1.58. Assuming that the errors are normally distributed, the conventional “95% confidence interval” can be calculated from p ± 1.96 S.E. or p ± 3%.
The fact that, on average, the actual monthly bills of responding firms exceeded those of the total original sample by $14/firm ($127–$113) indicates that nonresponse operated to yield a biased sample of responding firms, wherein heavy telephone users were overrepresented and lighter users were underrepresented. Comparing the reported and actual bills of responding firms, we further see that systematic measurement error was present, leading to overreporting by an average of $9/firm ($136–$127). Consider the consequences of relying on the billings reported by survey respondents to the mail survey to estimate the total sample’s true bill. The combined effects of sample bias due to nonresponse and overreporting due to positive measurement errors would have led to an estimate of mean bill size that was inflated by $23/firm ($136–$113) or about 20%.

Of course, it is unusual to have access to the “true” values of quantities estimated in surveys but comparisons like those discussed above serve as a sobering reminder of the challenge inherent in designing a survey. What can be done to minimize fallibility? Enlightenment can represent the first step down the path to improvement and a familiarity with the nature and sources of survey errors naturally leads to provocative questions about how to enhance the dependability of results. For example, in the above case, could nonresponse bias have been reduced by using a different method of data collection, such as face-to-face or telephone interviews? In a small business, who is likely to be most knowledgeable about the firm’s telephone expenses and thus best qualified to serve as a survey informant? Would measurement error be diminished if the questionnaire encouraged the informant to consult records? Would it be preferable to request information about telephone expenses in each of three specified months rather than asking for an “average for the past three months”? An understanding of possible errors will expand the range of options considered and often lead to adoption of methods that improve accuracy.

**Nonresponse Error**

Nonresponse error reduces the size of the sample available for analysis and constitutes a threat to the representativeness of the final sample. Since sampling error varies inversely with sample size, any loss of sample size due to nonresponse will increase the magnitude of random sampling error realized relative to the level anticipated if no allowance for nonresponse is made in planning the study.

More seriously, nonresponse is likely to cause systematic error or bias, as was the case in the telephone expense study discussed earlier. While the essence of a probability sample is the specification of procedures to ensure that elements of the population to be included in that sample are randomly drawn, selectivity invariably enters the picture once particular data collection methods are introduced to implement the overall sampling design. If respondents were themselves a random subsample of the total sample originally drawn, then nonresponse would not bias the final results. However, since responding to an interview or questionnaire is a conscious social act, it is unlikely to conform to the impartial laws of probability. Hence it is imperative to take steps to minimize both the nonresponse rate (i.e., the proportion of original sample that does not respond) and opportunities for uncontrolled selectivity to govern the capturing of study participants and their responses.

**Controlling access.** Clearly, availability and motivation are basic determinants of who responds to surveys and who doesn’t. Two classes of design variables allow the researcher to exert some control over availability or access. The first is choice of data collection method. The options are interviews (face-to-face or telephone) and self-administered questionnaires (delivered and retrieved via mail or personally). The differential effectiveness of these alternative modes of data collection in minimizing error depends on the subject matter of the survey and the population studied. No one mode of data collection dominates the others for all purposes and occasions and the choice involves

---

2Assael and Keon (1982) tested these alternatives and found that both personal and telephone interviews led to larger nonresponse biases than did the mail questionnaire.
tradeoffs. To illustrate, in a study of single parent households, telephone interviews may produce a higher response rate than a mail questionnaire, but if the subject matter is a socially sensitive one (e.g., controversial or threatening), then the mail questionnaire may offer respondents greater assurance of anonymity and thereby diminish selectivity bias.

Dillman (1978, chapter 2) has presented a useful framework for comparing personal interviews, telephone and mail, with respect to a set of highly relevant factors and the reader faced with designing a survey will find it helpful to apply Dillman’s framework to his or her problem.

A second means of influencing access relates to the administration of any chosen data collection method and involves decisions with respect to scheduling field work and providing for advanced notification and reminders or callbacks to sample members. These procedures can be successfully utilized to enhance response rates and a body of research bearing on their impact has accumulated. Reviews of relevant literature and data on response rates may be found in Steeh (1981) for personal interviews, Wiseman and McDonald (1979) for telephone surveys, and Fox, Crask, and Kim (1988) for mail surveys. Valuable information on scheduling face-to-face and telephone interviews is presented in Weeks et al. (1980, 1987).

Motivating response. Given access to an eligible sample member, then motivation to respond comes into play. A request to give an interview or to fill out a questionnaire represents intrusion into the life of the target individual. Why would or should one agree to participate and exert the effort to respond to questions conscientiously? The task of providing the requisite motivation places a heavy burden on what is communicated to respondents at the initial point of contact.

Dillman (1978, pp. 12–19) offers a perceptive analysis of respondent behavior as a special case of “social exchange.” Briefly, social psychological theories of exchange posit that an individual’s propensity to engage in a particular behavior is a function of the expected rewards and costs that person associates with doing so. This perspective leads Dillman to suggest that there are three things that must be done to maximize survey response: “minimize the costs of responding, maximize the rewards for doing so, and establish trust that those rewards will be delivered” (p. 12, emphasis added). Dillman’s suggestions for accomplishing these directives bear repeating:

1. Reward respondents by:
   a) showing positive regard,
   b) giving verbal appreciation,
   c) using a consulting approach,
   d) supporting his or her values,
   e) offering tangible rewards,
   f) making the questionnaire interesting.

2. Reduce costs to the respondents by:
   a) making the task appear brief,
   b) reducing the physical and mental effort required,
   c) eliminating chances for embarrassment,
   d) eliminating any implication of subordination,
   e) eliminating any direct monetary costs.

3. Establish trust by:
   a) providing a token of appreciation in advance,
   b) identifying with a known organization that has legitimacy,
   c) building on other exchange relationships.
Measurement Error

Thus far, we have focused on mean levels, computed by aggregating and averaging across the relevant data for each member of the sample. Thus, in the telephone example, on average, small businesses overstated their monthly telephone bills by $14 in responding to a mail survey. We would, of course, expect some variability among individual firms with respect to not only the magnitude of their reporting errors, but also their direction—most overreporting but some underreporting, and perhaps a few being perfectly accurate. Hence, we need to recognize the possible presence of both systematic and random error components in responses to a measuring instrument.

The distinction between systematic and random measurement errors is closely related to two important properties of measurement instruments: reliability and validity. Reliability is concerned with random measurement error and relates to questions about how stable, consistent, or reproducible are the responses obtained from some measurement instrument. Validity, as the term is applied to measures, involves questions about whether an instrument or procedure measures the concept that it claims to measure and, if so, how adequately it does so. What do SAT scores or GMAT scores or Nielsen television program ratings really measure? Do copy tests measure advertising effectiveness? What are meaningful indicators of corporate excellence and innovativeness? The presence of both systematic and random measurement errors in responses threaten a measure’s validity. Hence, in evaluating a measure or choosing among procedures it is important to consider their comparative reliability and validity. A further discussion of these concepts may found in the appendix of this note.

What Data Should Be Sought?

The first step in developing a questionnaire is deciding what information you want to find out. The challenge for the person designing the questionnaire is to assume the role of an interpreter who must not only speak the “language” of both study sponsor and respondent but also understand their points of view. To accomplish this “translation,” you must first map research purposes and goals into data requirements, then define the role and tasks to be performed by survey participants in supplying those data, and finally formulate questions that will elicit the target reports and assessments.

Types of Roles and Data

The researcher begins by generating a list of the kinds of data and information that are needed to address the particular set of research questions and/or hypotheses motivating the study to be undertaken. As a step toward relating these data requirements to the frame of reference of someone serving as a participant or subject in the study, it is helpful to classify the items comprising the list according to the following matrix, which cross classifies three types of data which may be sought with two kinds of roles that a study participant may be asked to assume in providing those data.

<table>
<thead>
<tr>
<th>Role of Participant</th>
<th>Behavioral Reports</th>
<th>Cognitive Assessments</th>
<th>Affective Assessments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respondent</td>
<td>Actions</td>
<td>Knowledge</td>
<td>Preferences</td>
</tr>
<tr>
<td></td>
<td>Events</td>
<td>Opinions</td>
<td>Attitudes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Beliefs</td>
<td></td>
</tr>
<tr>
<td>Informant</td>
<td>Practices</td>
<td>Perceptions</td>
<td>Satisfactions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Judgments</td>
<td></td>
</tr>
</tbody>
</table>
When a question asks a participant to say something about herself or himself, that person is cast in the role of a respondent. Alternatively, questions may ask about other persons or social entities. Hence, when the subject of a question is someone other than the participant, such as one or more members of a family, group, organization, or community of which he or she is a member, then that participant is being asked to serve as an informant (Seidler 1974). We use the term reports to denote responses that pertain to behavior (i.e., past or current acts, events, and practices); we apply the label assessments to responses that relate to mental states: either cognitions (i.e., information, knowledge, beliefs, judgments, images, and perceptions) or affects (i.e., feelings, emotions, and satisfactions).

These distinctions between types of roles and data are important because they begin to clarify the nature of the study participant’s task. For example, the internal processing required to provide a self-report about one’s own behavior is quite different from that involved in acting as an informant about the practices of an organization. Accordingly, the nature and magnitude of memory errors that may afflict responses to these two types of questions are also likely to be dissimilar.

Task Considerations

In order for a questioning procedure to succeed, three conditions must be satisfied for each respondent/informant: (a) comprehension of his or her role and the nature of the information being sought, (b) accessibility of that information, and (c) motivation to assume the role and engage in the requested task behavior (Cannell and Kahn 1968). The ease or difficulty of meeting these conditions varies with the nature of the task required of survey subjects. Based on an extensive review of research on measurement errors in surveys, Sudman and Bradburn (1974, pp. 8–13) formulated the following hypotheses about how various dimensions of the respondent’s task contribute to the level of error in their responses:

1. The greater the degree of structure, the lower the magnitude of measurement errors.

2. The greater the problems of self-presentation elicited by a question (i.e., the less socially desirable some answers or responses are perceived to be), the greater the pressure on the respondent to answer a question, or the more controversial the subject matter, the greater the magnitude of measurement errors.

3. The greater the saliency of the information sought, the lower the magnitude of measurement errors.

The problems of low saliency and self-presentation will vary according to the topic under study and the population surveyed; what can be done to ameliorate their effects may be limited. For some of these problems, however, special procedures have been designed that may be useful. For example, Sudman and Bradburn (1974, pp. 9 and 68–84) point out that in the case reporting behavior, low saliency is an issue primarily because of memory errors that may be reduced by employing procedures that aid or bound recall by providing specific cues such as lists, pictures, and dates and encourage subjects to consult records. Similarly, special methods of asking survey questions on controversial or socially sensitive topics have been devised (Bradburn and Sudman 1979) such as the “randomized response technique,” which preserves the confidentiality of individual responses (Fox and Tracy 1986). Indirect and unobtrusive methods of measuring attitudes (e.g., projective techniques and the lost letter procedure) may also be considered (Kidder and Campbell 1970; Webb, et al., 1981).

Another challenge in structuring the participant’s task is deciding whether to ask a global (summary) question or a series of specific questions which decompose the overall task into component parts. This issue frequently arises in connection with the use of informants who are asked, as Seidler (1974, p. 817) has wryly noted, “at least implicitly, to perform calculations otherwise left to the computer.” Specificity is favored over generality on the grounds that it facilitates control over task
interpretation and processing. Specific questions may be especially helpful when asked to assess such things as the price and quality of one firm’s products compared with competitors’ (Phillips, Chang, and Buzzell 1983), patterns of influence in organizational decision making (Silk and Kalwani 1982), and control in customer-supplier relations (Phillips 1981).

When one uses general questions, supplying informants with detailed instructions about what data or judgments are relevant and how calculations should be made simplifies the task and reduces the possibility of measurement error.

The questionnaire designer needs to be diligent in monitoring these matters and heed the warning of Cannell, Oksenberg, and Converse (1977, p. 309) who observed: “The demands placed on the respondent by many survey questions are greater than generally has been realized, and the respondent’s inability or unwillingness to meet these demands is a major source of invalidity.”

Questionnaire Design

Sources of Design Guidelines and Advice

In designing a question, one must address three basic issues: (1) What form of question should be employed, (2) how the question should be worded, and (3) whether response alternatives should be presented, and if so, which ones and in what format? A variety of solutions are available for each of these design problems. The choices should be made carefully since survey results can be sensitive to the form and wording of questions and response categories. Hence, the questionnaire designer should be familiar with the options available and guidelines for choosing among them.

Considerable literature has accumulated on the subject of questionnaire design and several readable volumes exist which are rich sources of examples and useful guidance. The following are some reference works which can be recommended to the novice questionnaire designer:


The prescriptions offered in these works are based on a modest amount of scientific evidence combined with large amounts of experience and common sense. Converse and Presser provide the most compact treatment (75 pages and available in paperback) but their text does not deal explicitly with self-administered questionnaires. Dillman treats both mail and telephone surveys in a comprehensive and detailed manner and includes a useful framework for choosing among face-to-face, telephone, and mail surveys. Payne’s is the classic work on questionnaire wording, a storehouse of examples and thoughtful counsel that is also enjoyable to read. Sudman and Bradburn offer balanced and comprehensive coverage of all three methods of data collection. Other relevant reviews are found in Dijkstra and van der Zouwen (1982) and Hippler et al. (1987).

The impact of small changes in question wording has often been demonstrated in simple experiments known as “split ballot” studies, in which different versions of a question are administered to randomly selected halves of the same sample. Evidence from such studies can help in
identifying sources of measurement errors and in suggesting ways to avoid or reduce them. Consider the following example from a routine consumer survey carried out for a chemical firm.

Respondents in a nationwide probability sample of 1,000 adults were asked to agree or disagree with a series of opinion statements about wash-and-wear clothing (O’Neill 1967). The sample of respondents was randomly split, and one half was presented with a “positively” worded version of an item while the other half was given a “negatively” worded counterpart. Below are the distributions of responses observed for two of the questions—shown as percentages of each split sample (n=500) who expressed a favorable, unfavorable, or neutral opinion about wash and wear clothing.

<table>
<thead>
<tr>
<th>Item</th>
<th>Wording*</th>
<th>Favor.</th>
<th>Unfav.</th>
<th>Neut.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Clothes like these (cut/increase) the cost of keeping clothes clean.</td>
<td>Pos.</td>
<td>79</td>
<td>12</td>
<td>9</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Neg.</td>
<td>76</td>
<td>13</td>
<td>11</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Diff.</td>
<td>+3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Clothes like these are (easy/hard) to get dirt out with washing.</td>
<td>Pos.</td>
<td>88</td>
<td>6</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Neg.</td>
<td>76</td>
<td>12</td>
<td>12</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Diff.</td>
<td>+12</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Data reported in O’Neill (1967, pp. 99–100).

* The alternative wordings used in the two versions of each question are shown in parentheses—the first word represents the “positive” wording and the second, the “negative” wording. Note that “disagreeing” with a “negatively” worded statement is treated as a “favorable” response.

Note that for both items, favorable opinions were more likely to be expressed for positively worded versions of the statements than for the negatively worded alternative. The standard error of the difference between estimates of a proportion for two samples, each consisting of 500 respondents, is about 3%. Thus, the difference of 3 percentage points in Item #1 is within the range of differences to be expected due to random sampling error. However, the difference of 12 percentage points in Item #2 is much larger than that which can be confidently attributed to sampling fluctuations. This difference is an example of the type of measurement error or response effect that results from unbalanced questions—questions that present only one side of an issue or opinion (Schuman and Presser 1981, chapter 7).

In light of such findings, generally accepted survey practice favors balanced questions (e.g., Payne 1951, chapter 4). In some instances, balance may be achieved in a straightforward manner—e.g., “Do you support or oppose passage of the Equal Rights Amendment?” Phrasing the question about wash-and-wear clothing in this manner would be more difficult, however. A construction such as, “Do wash and wear clothes increase or decrease cleaning costs?” poses another difficulty, that of an “implied alternative” (Payne 1951, p. 56). A better formulation would be: “Compared with wool, do wash and wear clothes increase or decrease cleaning costs?”

In addition to applying the kinds of general guidelines and advice found in the recommended references and illustrated above, the questionnaire designer should also seek out expertise bearing on the particular substantive domain to be investigated. Strongly recommended are informal interviews and/or focus groups with a small number of respondents from the target population. This can help compensate for whatever social distance may exist between the questionnaire designer and his/her informants or respondents and thereby alert him/her to problems of unfamiliar language and concepts, low salience, or threatening subject matter. In many fields, specialized measurement instruments have been developed, and their reliability and/or validity tested. Such material may be uncovered through a literature search or discussions with people who...
have conducted previous studies in the same area. Much can also be learned from examining a questionnaire that has already been used in similar work.

**Questionnaire Organization**

The next task is to arrange the sequence of questions so as to form an overall questionnaire. Dillman (1978, pp. 123–127) has suggested four principles for deciding how questions should be ordered.

1. Order questions along a descending gradient of social usefulness or importance; those that the respondent/informant is most likely to see as useful come first and those least useful come last.

2. Group questions that are similar in content together, and within content areas, by type of question (e.g., reports of behavior versus assessments of attitudes).

3. Take advantage of cognitive ties that subjects are likely to make among groups of questions to create a sense of flow and continuity.

4. Within any topic area, position questions that are least likely to be objectionable before the more objectionable ones.

Application of these principles often involves important judgments and compromises which should be re-examined in light of experience gained from pretesting. Questions used to obtain information on demographic characteristics and other background or classification variables are usually placed at the end of the questionnaire. In designing such questions, keep in mind anticipated comparisons of the study’s results with other studies or sources of information. For surveys among the general population, certain standard demographic categories (e.g., age, education, income) have been developed and are presented in Sudman and Bradburn (1982, chapter 7).

Two other tasks must be completed. First, cover letters, introductions, and instructions need to be prepared. Dillman’s (1978, chapters 5 and 7) discussion of these topics is highly recommended. It is important to address growing public and government concern over privacy and confidentiality issues in relation to the collection and uses of survey data (National Research Council 1979). Promises of confidentiality made to respondents and informants are a responsibility that the researcher must take very seriously. Second the questionnaire must be formatted and precoded to facilitate data processing. Dillman (1978) and Sudman and Bradburn (1982) provide detailed instructions on these matters.

**Critical Reviews and Pretests**

Critical review and pretesting of the questionnaire are vital. More than one round of review and testing may be necessary in order to develop a questionnaire that is sufficiently refined to warrant asking respondents to make a serious commitment of their time to answering it. As Sudman and Bradburn (1982, p. 283) advise “If you do not have the resources to pilot test your questionnaire, don’t do the study.”

Reviews and tests should help identify excesses and omissions. Virtually all discussions of questionnaire development admonish the designer to ask, “Is this question really necessary or merely interesting?” and urge, “Keep it short and simple.” One useful exercise is to go through a draft of the questionnaire item by item and simulate the outcome of the study by creating dummy tables indicating how the data will be eventually analyzed and reported. Such a disciplined effort invariably pays off by uncovering both excesses and omissions.
A tedious but nonetheless effective way to detect wording problems is to check each questionnaire item against the following list of questions proposed by Dillman (1978, pp. 97–118):

1. Will the words be uniformly understood?
2. Does the question contain abbreviations or unconventional phrases?
3. Is the question too vague?
4. Is the question too precise?
5. Is the question biased?
6. Is the question objectionable?
7. Is the question too demanding?
8. Is it a double question?
9. Does the question have a double negative?
10. Are the answers mutually exclusive?
11. Does the question assume too much about what respondents know?
12. Is the question technically accurate?
13. Is an appropriate time referent provided?
14. Can the question be understood when taken out of order or context?
15. Can responses be compared to existing information?

Summary

Questionnaire designers should think of response behavior as a special form of social exchange. The challenge is to construct a questionnaire that promotes access, comprehension, and motivation among participants. Roles and tasks must be carefully defined and structured with the goal of establishing trust, minimizing costs, and maximizing rewards associated with complying with the intrusive request of responding to a questionnaire. Much can be done to reduce nonsampling errors and enhance the ultimate value of a survey by following an orderly, sequential approach to the development of measuring instruments. Such a process should make explicit provisions for obtaining early feedback via pretesting and checking the instrument for consistency with relevant design guidelines.
Appendix
Measurement Errors, Reliability, and Validity

A Model of Measurement Error

To lay the groundwork for a discussion of the concepts of reliability and validity, we begin with a simple model that decomposes responses to an item or question into three components:

\[ X_i = S_i + T_i + e_i \]  

where \( i \) is an index that identifies a particular respondent or informant.

The above measurement model assumes that an individual’s observed response is the sum of his/her “true” score (\( T_i \)) plus two error components, one systematic or fixed (\( S_i \)) and the other random or transitory (\( e_i \)). To illustrate the distinction, imagine a small business manager responding to the same question about his/her firm’s telephone expense on two separate occasions, where the firm’s actual bill is $130. Relying on memory, assume he/she reports $90 on the first occasion and $110 on the second for an average of $100 across the two reports. Now suppose that on both occasions, the manager overlooked the expense of a special line for a fax machine that cost $30. The latter omission, consistently made, would represent a systematic error or bias. The variations between occasions of plus or minus $10 around the mean of $100 could be regarded as a random component of measurement error inasmuch as sometimes it is “high” and other times it is “low,” but across occasions one will tend to “cancel out” the other.

Reliability

We may use the measurement model represented by equation (1) above to examine two important properties of measurements: reliability and validity. Suppose we administered some measuring instrument to all members of some population and computed the mean and variance of their observed responses. If we assume that the true scores and the systematic and random error components are mutually independent (e.g., the magnitude of true scores and error components are unrelated), then it can be shown that:

\[ U_x = U_s + U_t \]  
\[ \text{var} (x) = \text{var}(s) + \text{var}(t) + \text{var}(e) \]

where \( U_x, U_s, \) and \( U_t \) are the expected population values of the means for the observed scores, systematic errors, and true scores, respectively and the \( \text{var}(\ ) \) terms represent the population variances of the observed scores (\( x \)), systematic errors (\( s \)), true scores (\( t \)), and random errors (\( e \)).

Note that we are defining random measurement error (\( e \)) such that the expected population value of its mean is zero (i.e., \( U_e = 0 \)) and hence it does not affect the expected value of the mean of the observed scores (\( U_x \)), which accounts for its absence in equation (2). However, random measurement error does add a component, \( \text{var}(e) \), to the variance of the observed scores, \( \text{var}(x) \), as indicated by equation (3). Thus, we can see from (2) and (3) the adverse effects that both systematic and random measurement errors have on our ability to make inferences about the values of population parameters which we may seek to estimate from our fallible survey responses in order to address the goals of the study. Not only does systematic error bias the value of the mean of the observed scores as an estimate of the true score mean, but the precision of our estimates is also diminished. That is, since the
variance of the observed scores is inflated by the presence of systematic and random error components, the standard errors of other statistics we may compute, such as that for the mean, will also be increased, as compared with what their magnitude would be without such errors.

We are now in a position to understand the concept of reliability and see why knowledge of this property of a measuring instrument is important. Conceptually, reliability is defined as the ratio of true score variance to observed score variance:

\[ R = \frac{\text{Var}(t)}{\text{Var}(x)} \]  

where \( R \) denotes the index of reliability, whose value may vary between zero and one.

Reliability answers the question: how much of the variability in the observed scores can be attributed to “true” differences among members of the population as opposed to differences resulting from systematic and random measurement errors? Thus, the higher the reliability, the greater the proportion of observed score variance that reflects true score variance.

The most common method of assessing reliability is to administer the same measuring instrument to the same set of respondents on two different occasions and correlate the two sets of observed scores. This correlation is known as a “test-retest reliability” coefficient. The rationale behind the method relates to the intuitive interpretation of reliability as “stability” or “reproducibility.” Applying the measurement model represented in equation (1) to both test and retest occasions, we can see that if the true score and systematic error components remain fixed, then any variation in observed scores would be attributable to random error variation—i.e., transient factors or “noise.” Therefore, the greater the test-retest correlation of observed scores, the greater their stability.\(^3\)

It should be noted, however, that if the observed scores contain a stable component of systematic error, then such a test-retest correlation will represent an index of the stability of the observed scores but will not conform to the definition of reliability given in equation (4). In that case, it will represent a ratio of the sum of the systematic and true score variances to observed score variance rather than the ratio of true score variance to observed score variance. Of course, if the observed scores are free of systematic errors (i.e., \( \text{var}(s)=0 \)), then a test-retest correlation will correspond to the definition of reliability given in equation (4). For a discussion of other methods of assessing reliability, the interested reader is referred to Bohrnstedt (1983).

**Validity**

In addition to reliability, it is also desirable to demonstrate the validity of a measure. Validity is a multifaceted concept, and a variety of empirical methods are employed to assess it. We briefly examine some of the more frequently encountered approaches.

*Content validity* is concerned with representativeness. Does a measure adequately represent the substance of the property being measured? Consider the operational definition of unemployment used by the Bureau of Labor Statistics (BLS) in its monthly surveys to estimate this key economic indicator. A respondent who has not worked during the week of the survey but who is available to do so is not counted as “unemployed” unless he/she has looked for work during the preceding four week period. As Norwood (1988, p. 285) observes, “The line that divides active job seekers from those

\(^3\)The interpretation assumes that the random measurement errors present in the test and retest scores are statistically independent but have the same variance.
not in the labor force at all is really difficult to draw with precision and finality.” Not surprisingly, the measurement of unemployment has long been controversial.

Content validity is typically assessed by eliciting expert judgments. Thus, advisory committees of labor market analysts are periodically convened to review BLS methods and recommend modifications. Similar approaches are used to develop psychological tests and may be applied in other domains as well. For further details, see Bohrnstedt (1983).

Criterion-oriented validity is an issue frequently raised in applied research. Here the focal measure is related to one or more external variables or criteria, selected on the basis of a priori reasoning. An important distinction is whether the relationship investigated is a concurrent one or a predictive one. An example of concurrent validity is provided by Haley and Case (1979), who examined the association between various measures of brand attitudes and reports of current brand usage obtained in the same interview. On the other hand, Kalwani and Silk (1982) analyzed the predictive validity of buying intention scales by relating such measures taken at one point in time to observations of purchasing behavior obtained subsequently from the same sample of respondents. Predictive relations generally provide a more stringent validation test than concurrent ones but the former may require complex modeling to isolate the linkages between the focal and criterion variables from exogenous factors and confounds.

Construct validity is concerned with the question of whether a particular instrument or method measures what it purports to measure. Do counts of patents and scientific publications measure technological innovativeness (Narin and Frame 1989)? Can both peer and supervisor ratings be used to assess managerial performance (Lawler 1967)?

Construct validity is assessed by means of tests for convergence and discrimination among alternative measures (Campbell and Fiske 1959). Convergent validity is indicated by the extent of agreement between two separate methods of measuring the same construct. But a measure may be invalidated by a lack of discrimination, if, for example, it is found to correlate too highly with another measure from which it was intended to differ. An example from advertising research illustrates this point. Appel and Blum (1961) showed that a particular method of measuring advertisement readership produced scores for ads before they were run that correlated strongly with those obtained after the ads had actually appeared. This finding indicated that the method lacked the ability to discriminate between false and true claims of readership. The correlation that was observed was attributed to a stable component of erroneous readership present in both administrations of the method.

The issue of discriminant validity is especially germane when the same or similar methods are used to measure different constructs. For example, respondents may be asked to rate several options on the same set of scales (e.g., all employing the same response format) administered on the same occasion. Such practices run the risk that a shared component of systemic error related to scale format will contaminate the entire set of ratings and make it difficult to detect the presence of true differences in the ratings of the alternatives.

Tests for convergent and discriminant validity require that several constructs be measured by each of several methods (Campbell and Fiske 1959, Bagozzi 1980). Such assessments have been performed for a number of constructs used in diverse areas of management research, including employee performance appraisal (Lawler 1967), customer-supplier relations (Phillips 1981), patterns of influence in organizational purchase decision making (Silk and Kalwani 1982), and business strategic orientation (Venkatraman 1989).
References


