United States General Accounting Office

GAO

Program Evaluation and Methodology Division

October 1993

Developing and Using Questionnaires

GAO/PEMD-10.1.7

Preface

GAO assists congressional decisionmakers in their decisionmaking process by furnishing analytical information on issues and options under consideration. Many diverse methodologies are needed to develop sound and timely answers to the questions that are posed by the Congress. To provide GAO evaluators with basic information about the more commonly used methodologies, GAO's policy guidance includes documents such as methodology transfer papers and technical guidelines.

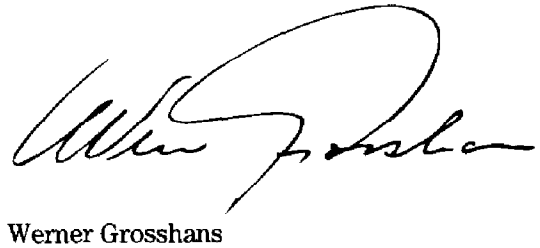
The purpose of this methodology transfer paper is to provide evaluators with a background that is of sufficient depth to use questionnaires in their evaluations. Specifically, this paper provides rationales for determining when questionnaires should be used to accomplish assignment objectives. It also describes how to plan, design, and use a questionnaire in conducting a population survey. We do not expect GAO evaluators to become experts after reading this paper. But we do hope that they will become familiar enough with questionnaire design guidelines to plan and use a questionnaire; to make preliminary designs and assist in many development and testing tasks; to communicate the questionnaire requirements to the measurement, sampling, and statistical analysis experts; and to ensure the quality of the final questionnaire and the resulting data collection.

The present document is a revision. An earlier version was authored by Brian Keenan and Marilyn Mauch in 1986. This revision, authored by Brian Keenan, includes new material on cognition as well as on a number of developments in pretesting that have occurred since then. As such, the present document supersedes the 1986 version.

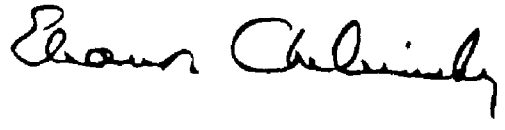
Developing and Using Questionnaires is one of a series of papers prepared and issued by the Program

Evaluation and Methodology Division (PEMD). The purpose of the series is to provide GAO evaluators with guides to various aspects of audit and evaluation methodology, to illustrate applications, and to indicate where more detailed information is available.

We look forward to receiving comments from the readers of this paper. They should be addressed to Eleanor Chelimsky at 202-512-2900.



Werner Grosshans
Assistant Comptroller General
Office of Policy



Eleanor Chelimsky
Assistant Comptroller General
for Program Evaluation and
Methodology

Contents

Chapter 1

Using Questionnaires

This paper describes how to design and use questionnaires. Such information is important for GAO evaluators for two reasons. First, GAO frequently uses questionnaires to collect data. Second, the questionnaire is a method with a high potential for error if not designed and used properly.

GAO employs questionnaires to ask people for figures, statistics, amounts, and other facts. We ask them to describe conditions and procedures that affect the work, organizations, and systems with which they are involved, and we ask for their judgments and views about processes, performance, adequacy, efficiency, and effectiveness. We ask people to report past events and to make forecasts, to tell us about their attitudes and opinions, and to describe their behavior and the behavior of others.

Questionnaires are popular because they can be a relatively inexpensive way of getting people to provide information. But because they rely on people to provide answers, a benefit-risk consideration is associated with their use. People with the ability to observe, select, acquire, process, evaluate, interpret, store, retrieve, and report can be a valuable and versatile source of information under the right circumstances. However, the human mind is a very complex and vulnerable observation instrument. And if we do not ask the right people the right questions in the right way, we will not get high-quality answers.

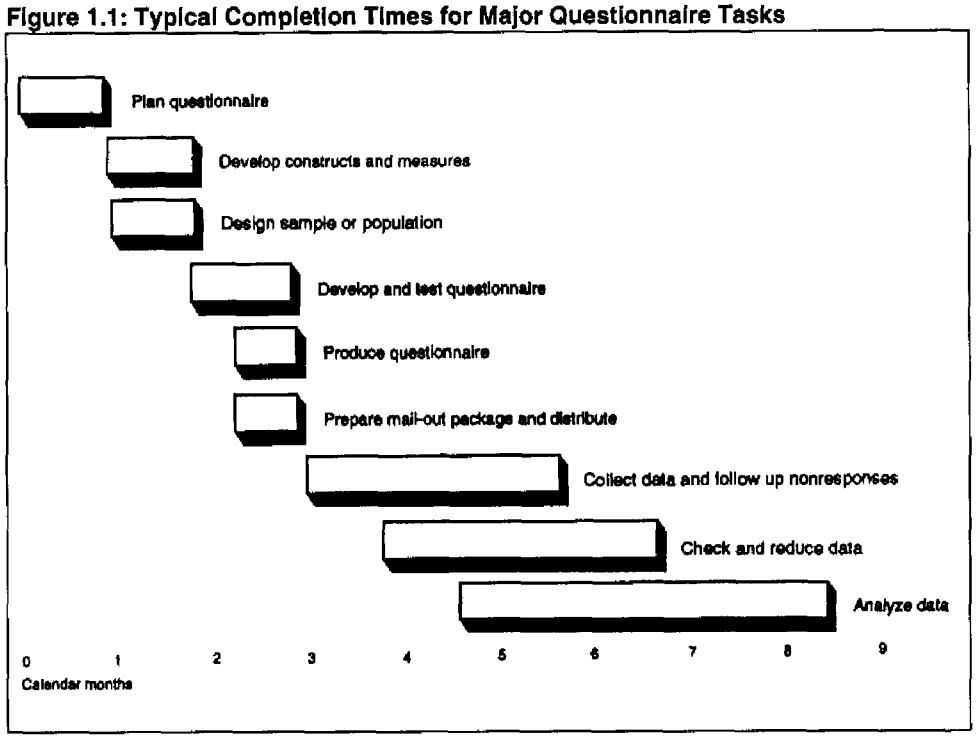
This holds true for even the simplest of questions. An easy way to demonstrate this is to do a simple straw poll, like asking co-workers how they came to work. One may answer "By way of New York Avenue" or give some other route description. Another answer to the same question may be "by car pool." If you continued this straw poll, many of the answers would be unusable if your intent was to learn modes of transportation to work.

Asking good questions in the right way—the focus of this paper—is both a science and an art. It is a science in that it uses many scientific principles developed from various fields of applied psychology, sociology, cognitive research, and evaluation research. It is an art because it requires clear and interesting writing and the ability to trade off or accommodate many competing requirements. For example, a precisely worded, well-qualified, unambiguous question may be stilted and hard to read. Questions must be clear, interesting, and easy to understand and answer. In addition to asking the right questions, evaluators need to be aware of other principles dealing with questionnaire design and administration that are also covered in this paper.

Overview of Tasks in Using Questionnaires

Using even a simple questionnaire is not always simple. Numerous major tasks to develop and use a questionnaire must be completed in a logical sequence. After deciding to use a questionnaire, evaluators must plan the questionnaire, develop measures, design the sample, develop and test the questionnaire, produce the questionnaire, prepare and distribute the mailout or interview packages, collect the data and follow up with nonrespondents, perform checks to ensure the quality of responses, and reduce and analyze the data. Figure 1.1 reviews these major tasks. Except for the data collection, these processes are very similar regardless of whether the questionnaire is to be designed for the mail or a telephone or face-to-face interview. When interviewers are used, however, they must also be trained, which adds another major task.

Figure 1.1: Typical Completion Times for Major Questionnaire Tasks



After describing important factors to consider when deciding to use a questionnaire, we briefly cover, in the remaining sections of this chapter, the major tasks listed in figure 1.1 and refer to subsequent chapters that provide detailed instructions. We do this to give an overview of the scope of work required to plan, develop, and implement a questionnaire and to show what the reader can expect to find in each of the subsequent chapters. Overall, the organization of this paper parallels the logical sequence of tasks undertaken when developing and using questionnaires.

Deciding to Use Structured Questionnaires

One of the first decisions evaluators have to make is whether to use a questionnaire or some other method to collect the data for the job. In many situations, other data collection techniques may be superior. In fact, over the past years other techniques were recommended by technical design teams for about one of every three proposed GAO questionnaires. The decision to use questionnaires should be made only after carefully considering the comparative advantages and disadvantages of the various ways of administering questionnaires over other data collection techniques.

Data Considerations

Data can be collected in a variety of ways, such as field observations, reviews of records or published reports, interviews and standardized mail, and face-to-face or telephone questionnaires. The selection of one technique over another involves trade-offs between staff requirements, costs, time constraints, and—most importantly—the depth and type of information needed. For example, if the objective of the assignment is to determine the average per acre charge and the income derived from public grazing-land permit fees, the evaluator might consider using structured data collection forms or pro forma work papers to manually retrieve data from the case files in record storage. However, if the objective is to determine how much land the ranchers are willing to lease and how much per acre they are willing to pay, a mail, telephone, or face-to-face survey of ranchers would be necessary.

Questionnaires are frequently used with sample survey strategies to answer descriptive and normative audit or evaluation questions. They are often less central in studies answering impact, or cause-and-effect, questions. While operational audits and impact, or cause-and-effect, studies are often not

large-scale efforts, questionnaires can be used to confirm or expand their scope.

Questionnaires can be useful when the evaluator needs a cost-effective way to collect a large amount of standardized information, when the information to be collected varies in complexity, when a large number of respondents are needed, when different populations are involved, and when the people in those populations are in widely separated locations.

Furthermore, questionnaires are usually more versatile than other methods. They can be used to collect more types of information from a wider variety of sources than other methods because they use people, who can report facts, figures, amounts, statistics, dates, attitudes, opinions, experiences, events, assessments, and judgments during a single contact. People can answer for a specific type of source, such as members of a health maintenance organization, or for a variety of types of sources, such as local, state, and federal government officials.

Questionnaires are difficult to use if the respondent population cannot be readily identified or if the information being sought is not widely distributed among the population of those who hold the knowledge. Furthermore, questionnaires should not be used if the respondents are likely to be unable or unwilling to answer or to provide accurate and unbiased answers or if the questions are inappropriate or compromising.

In general, questionnaires should not be used to gather information that taxes the limitations of the respondent. Sometimes people are not knowledgeable or accurate reporters of certain kinds of information. They remember recent events much better than long-past events. They remember salient and routine events and meaningful facts but do not remember

details, dates, and incidental events very well. For example, veterans might accurately report that doctors made medical examinations for Agent Orange effects on their eyes, ears, nose, throat, genitals, and pelvis but might substantially underreport skin examinations. If the information were needed on skin examinations, other sources, such as medical records, might be more useful. However, there are exceptions, particularly when the respondents are highly motivated.

Structured questionnaires are also not particularly well suited for broad, global, or exploratory questions. Because respondents have many different forms of reference, levels of knowledge, and question interpretations, the structured methodology limits the evaluators' ability to vary the focus, scope, depth, and direction of the line of inquiry. Such flexibility is necessary to accommodate variances in the respondents' perceptions and understanding that result from such questions.

Most of the people from whom GAO evaluators seek information are members of special populations, such as federal and state government employees, welfare recipients, or company executives. Unlike pollsters and market researchers, GAO evaluators rarely do a national population survey. Consequently, some of the mass survey techniques like random-digit dialing seldom apply to GAO work.[1] Also, GAO evaluators very rarely go back to the same population, and when they do, the time periods between surveys are so long that they usually have to redocument the population.

[1]Random-digit dialing refers to a telephone interview method that contacts people by dialing numbers at random. In some situations, usually when the population is hidden or not easily identified (for example, heads of households older than 65), this method may provide better access than other methods.

examine heads of households older than 65). this method may

Administration Considerations

If after considering the pros and cons of using questionnaires, a questionnaire is still the method of choice for data collection, the evaluators need to consider the most appropriate method of administration. The appropriateness of the method of administration—whether it be mail, face-to-face interview, or telephone—varies with the resources and constraints of the job, the abilities and motivation of the respondent population, and the requirements of the evaluation. All three methods have comparative advantages and disadvantages, depending on the time and cost constraints of the job, the characteristics of the respondent population, and the nature of the inquiry.

Mail questionnaires are usually more cost effective but require longer time periods than personal or telephone interviews. While mail questionnaires usually have higher development costs than telephone or face-to-face interviews, this is generally offset by the relatively inexpensive data collection costs. Mail questionnaires are the least labor intensive of the alternatives, with the labor costs limited to the effort needed to mail the questionnaire and track, follow up on, and edit the returns. Generally staff can mail hundreds of letters or edit scores of returns in a given day. Workers are not so productive with telephone and face-to-face interviews.

Because of the difficulty in establishing telephone or personal interview contacts and the one-on-one nature of interviews, these alternatives require more staff time. Interviewers usually do not complete more than 10 or 12 telephone interviews or two or three face-to-face interviews in a day. Furthermore, the travel requirements for personal interviews can be very expensive when compared to postage or telephone charges.

But mail questionnaires take longer to design and require longer periods for collecting and editing data than other choices. Extra care must be taken with the mail questionnaires because, unlike the other choices, there is no interviewer to help the respondent. Also, mail is a slow means of transmission, and mail questionnaires take two or three follow-ups. In summary, if money is tight and the subject matter can be phrased intelligibly for the respondent population, use the mail; if time is tight and staff time is not, use the face-to-face or telephone interview methods.

In addition to subject matter, respondent characteristics play a key role in the method of choice. For example, if the respondents are motivated and literate and have normal vision, the mail is often the best option; otherwise, use the telephone or an interviewer. If respondents cannot be readily located by address or telephone number but gather at particular places (such as restaurants, parks, or hospitals), then a face-to-face interview is the only option.

If the contact people are likely to conceal the identity of the intended respondent, and this is likely to make a difference, or if the evaluator is not sure that the intended respondent will get the questionnaire, then personal contact is better than telephone and telephone is better than mail. Also, if the respondent has a vested interest in giving biased reports that can readily be verified by inspection, then the face-to-face interview is the obvious choice.

However, if the contact has a likely chance of temporarily inconveniencing the respondent or the respondent has privacy concerns, then a mail survey has the advantage over the remaining choices.

Questionnaire characteristics also determine choice. Long, complex questionnaires designed to be

answered by simple checks or short fill-in-the-blanks are better suited for self-administered questionnaires than the interview method. However, the converse is often true if the questions require the composition of responses that are other than very short answers (most people would rather speak than write). Also, if the questionnaire has many complex and confusing skips that frequently require respondents to answer some questions but not others, then one of the interview methods is preferable to a mail or self-administered questionnaire.

In summary, evaluators should review the conditions and requirements of the data collection before deciding to use questionnaires and again before deciding the methods for administering the questionnaire. Mail questionnaires are a versatile, low-cost method of collecting detailed data. They are particularly adaptable to survey methods when the population is big, difficult to contact, likely to be inconvenienced, concerned about privacy, and widely dispersed. But mail questionnaires usually have a long turnaround time. The evaluators must be willing to invest the time required to carefully craft and test these questions. And the respondent must be willing and able and sufficiently literate and unbiased to accurately answer the queries. Interview methods, while much more expensive and more prone to bias, help insure against respondent error, have less turnaround time if sufficient staff is provided, and can be used to provide some interviewer verifications.

Planning the Questionnaire

Once evaluators decide to use a questionnaire, planning starts with this paper, which provides information on the procedures necessary to do each of the major tasks to design and use questionnaires. The next step is to review the evaluation design and audit plan and then mentally walk the job through each procedure necessary to design and implement a

questionnaire: developing the measures, designing the sample, developing and testing the questionnaire, producing the questionnaire, preparing the mailout or interview materials, and conducting the data collection, reduction, and analysis. A write-up of this mental walk-through, evaluated for comprehensiveness and feasibility, can serve as a basis for writing the implementation plan.

Developing the Measures

As evaluators do their planning, they will find that the scope of the effort is greatly influenced by information developed in the next two tasks—developing the measures and the sample design to ensure that the right questions are being asked of the right people. Remember that the questionnaire is an instrument used to take measures. To be sure it can do this, evaluators must first identify all the variables or conditions, criteria, causes, and effects that they want to measure. Next, evaluators analyze these variables and describe them so scientifically and precisely that they can be qualified, quantified, manipulated, and related. As explained in chapter 2, "Developing the Measures to Get the Questions," these measures define the requirements for the questionnaire. Questionnaires are designed by establishing a framework and sets of related questions that provide these measures.

Designing the Sample

Questionnaires are a way of asking the right people to take the measures needed to complete an evaluation. Before evaluators begin to write a question, it makes good sense to be sure they can find the people. The right people are representatives of a population who share the experiences the evaluators are interested in and who have or can get, will get, and will give them the information they need. Furthermore, evaluators must select these people scientifically, so the population these people represent can be talked

about rather than just the individuals contacted. This is called a population survey, and how to do a population survey with questionnaires is explained in chapter 3, "Designing the Sample or Population for Data Collection."

Developing and Testing the Questionnaire

Once the evaluators have established what to measure and who to ask to take the measures, they are ready to ask people to take these measures. Asking questions in the right way requires the evaluators to write sets of questions so that the answerer can easily understand precisely what information must be provided and, with little or no error, can easily provide this information. This means writing questions in a way that facilitates rather than interferes with the respondents' ability to understand the question and report the answer to the best of their ability. This simply stated task is deceptively complicated. To write good questions, evaluators must first understand something about the very complicated mental or cognitive process people use to answer questions. If evaluators access this cognitive process properly, the questionnaire can become a highly versatile and powerful instrument for observation and recall. If not, it can become a source of confusion and error.

The sets of inquiries or questions must then be organized into a draft instrument. This questionnaire is then tested, reviewed, and revised until it is proven that as an instrument it takes the required measures. Since completing these tasks is perhaps the most difficult part of the job and consumes the most resources, we devote nine chapters (chapters 4-12) to explaining some of the many known and tested ways to do this work.

Chapters 4-7 show how to facilitate the perception, acceptance, and understanding of the questions and

how to help respondents recall their mentally stored information. In chapter 4, "Formatting the Questions," we show how to present the question in the precise format best suited to get the specific type of information requested. We demonstrate what respondents are likely to consider as fair and unfair questions in chapter 5, "Avoiding Inappropriate Questions." In chapter 6, "Writing Clear Questions," we explain how to write a question that can be quickly, easily, and precisely understood by all respondents in the same way. And in chapter 7, "Developing Unscaled Response Lists," we explain how to write in a way that aids respondents as they cognitively search their minds to select the answers to questions.

Chapters 8 and 9 deal with the problem of bias and error. This problem has two sources: the question writer and the question answerer. Chapter 8, "Minimizing Question Bias and Memory Error," illustrates many of the typical mistakes question writers make and how to avoid them. Chapter 9, "Minimizing Respondent Bias," explains the ranges of capacities and limitations that respondents have in answering questions and how to make the most of the respondents' abilities and minimize the risk and compromise of their shortcomings.

Chapter 10, "Measurement Error and Measurement Scales in Brief," explains how to translate the question answers into qualitative and quantitative measures for use in GAO reports. Throughout chapter 10, we deal with how to write individual questions. However, when we put these individual questions together into a single questionnaire, they often interact with one another in a context that affects the measuring of the questions. Chapter 11, "Organizing the Line of Inquiry," shows how to organize these questions into a line of inquiry that can enhance the

quality of the answers and minimize unintended and interfering effects.

After finishing the first 11 chapters of this paper, evaluators should be able to help write the first draft of a questionnaire. But there is still much more to be done before evaluators can use this draft as a survey instrument. They should go through a quality-assurance procedure, which requires that the draft questionnaire be tested and validated. The methods for this task, and other quality assurance tasks carried out during data collection and analysis, are described in chapter 12, "Following Quality Assurance Procedures."

Producing the Questionnaire

Once the questionnaire has been tested and validated, and probably revised, the evaluators can put it in final form and use it to collect and analyze data to answer the assignment questions. Good questionnaires can be seriously compromised if they are not presented in a format that is easy to read and administer. Chapter 13, "Designing the Form and Layout," addresses this issue and shows the evaluator how to design the questionnaire type, format, and layout in a manner that greatly facilitates the user's ability to perceive and respond.

Preparing for and Collecting Data

Several administrative procedures, such as preparing the transmittal or contact letters or mail piece or interviewers' kits, must precede data collection. Data collection methods then involve such activities as mailing, contacting, interviewing, tracking, and following up on nonresponses. Poor quality in the execution of these fundamental and very important activities can cut the response rate by as much as 50 percent. To avoid this problem, we have documented procedures shown to be highly effective for mail surveys in chapter 14, "Preparing the Mail-out

Package and Collecting and Reducing the Data." Activities needed to check, edit, and prepare the data for computer processing are also covered in this chapter.

Analyzing Data

Chapter 15, "Analyzing Questionnaire Results," discusses some of the initial thinking and conceptualization that are important to the data analysis, including the development of a strategy and a plan for the data analysis. We do not describe data analysis methods since they are covered in Quantitative Data Analysis: An Introduction.[2] Chapter 15 concludes the discussion on using mail and self-administered questionnaires.

Telephone Surveys

Personal or telephone interviews are also important and useful methods for collecting structured data for GAO assignments. While the methodology for asking good questions developed in this paper applies regardless of whether the questions are asked in a self-administered mode, such as by mail, or in some other mode, such as a face-to-face or telephone interview, certain limitations are specific to each administration method. Those that apply to conducting telephone surveys are discussed in the concluding chapter 16, "Adaptations for the Design and Use of Telephone Surveys." Further details on personal interviews are presented in Using Structured Interviewing Techniques.[3]

[2]U.S. General Accounting Office, Quantitative Data Analysis: An Introduction, GAO/PEMD-10.1.11 (Washington, D.C.: June 1992).

[3]U.S. General Accounting Office, Using Structured Interviewing Techniques, GAO/PEMD-10.1.5 (Washington, D.C.: July 1991). Some information relevant to conducting face-to-face interviews is presented in chapter 12 of this paper, in a section dealing with pretesting techniques.

Chapter 2

Developing the Measures to Get the Questions

Deciding what and whom to ask appears to be a straightforward task. But appearances can be deceiving. And as we shall see in the next two chapters, this initial step must be thought through with careful consideration and structured to an elemental level of detail. The what and whom to ask decision lays the foundation for the focus and scope, the level of difficulty and complexity, the risk, completion times, data collection, analysis, and processing requirements and resources needed for the job. Hence, all the job plans are based on this decision. Furthermore, the three major sources of error—misspecification of variables, measurement error, and sampling error—are often introduced at this stage.

In this chapter, we discuss methods for documenting what a questionnaire should ask. This documentation will be used to develop a framework for writing the questions, describing the variables in scientific terms necessary for measurement, developing the measures, and specifying the variable relationships in order to check for misspecification of variable and measurement errors. In the next chapter, we discuss protocols for selecting the target population in ways that maintain the integrity of the design and minimize sampling error. Because deciding what to ask and deciding whom to ask it of are complex, we have described them in two chapters. However, in actual practice, deciding what and whom to ask go hand in hand and are among the few tasks in survey research that must be done interactively and iteratively. This is because the questions we ask are determined by both the need for information and the respondent's ability to provide this information.

To document the questionnaire framework, variable operationalizations, measures, and variable relationships, it is best to start with what we know about the requirements of the job and mentally work

in two directions, by thinking, first, in the abstract to integrate and conceptualize and, then, shifting to more concrete logic to define and analyze. At the start, evaluators usually find that some of the information they will need is very global, general, and abstract and other information is highly specific. However, most of the information they have gathered is at a middle level of detail, and they can begin by working with what they have. Information should be available from the job design, audit plan, evaluation framework, and previously gathered background material. Evaluators should conceptualize and organize this information into a framework of inquiry or types of questions that can be developed to yield answers to the evaluation questions. Often they may have to do additional research or additional thinking through to fill knowledge gaps.

Next, they must go in the other direction and think more concretely and analytically. They must specifically describe or operationalize these information requirements and develop measures that will satisfy these requirements. Finally, they should integrate these conceptualizations and analysis into a format that presents the key relationships of the measurement variables. The process needed to develop each document product is described in the following sections.

The Questionnaire Framework

Initially, the evaluators decide what constructs, traits, conditions, or variables are to be measured and how to measure them. The documentation for this task is sometimes referred to as a questionnaire framework. The framework is usually depicted as a taxonomical classification. It is a scheme that lays out the evaluation questions and all the information required to answer each question with ordered and specified relationships. In essence, the framework provides a

roadmap to identify and track the kind of data needed to answer the questionnaire.

A relatively uncomplicated example might be structured in response to the evaluation question "Is the size of the 4-year college associated with student performance?" The constructs (or the things evaluators want to measure) for college size and student performance and their relationships are identified for measurement development.

The identification of these constructs and their relationships influences the choice of data collection sources, methods, and measures. For instance, in the example above, we can readily see that there are alternatives: the use of extant data from various national graduate record achievement score data bases, surveys of administrative and academic deans, and so on. And just as the choice of methods and sources will force a choice of measures, so will the choice of measures determine the methods and sources.

Hence, these choices must be made interactively and iteratively. The relationship of college size to student performance was a simple example. In this case, evaluators might have been able to proceed without committing measurement considerations to paper, but it is nearly impossible to plan complex questionnaires without documentation. For example, consider the following evaluation question: "What are the needs of earth-orbiting satellite image users?" The answer to this question requires a plurality of complex considerations, constructs, and measures such as the identification of the different types of users (national and international scientists, political administrators, disaster managers, and earth resource managers) the identification of the national and international, geopolitical, and socioeconomic considerations that determine the type of use and the

measures of the quality of the information displays of the satellite and the relationships among the variables and constructs. This is a level of complexity that requires documentation.

As we can see by the example, the framework identifies, specifies, and justifies the need for the information, constructs, variables, measures, and variable relationships that the evaluator wishes to collect data on. It is a scheme for documenting the information needs requirements. It is not a questionnaire but rather the basis for the questionnaire.

Operationalizing the Constructs

So far we have talked in broad terms about ideas or concepts, traits or properties, and characteristics evaluators often like to measure—usually referred to as constructs. These constructs are not measures until the terms are specific enough to standardize. By "standardize," we mean that questions are designed and asked so that each recipient will understand and answer the same question in the same way. Different people reading the same questions need to have a common understanding. For example, one survey asked congresspersons about the "timeliness" of reports. Some respondents interpreted the construct "timeliness" as turnaround time while others interpreted it as getting the report information in time to use it for legislative decisions. As we can see, standardizing is very important because it enhances the objectivity of the resulting measure.

The first step toward standardization is to operationalize or to define the construct in concrete, specific, unambiguous, and contextual terms that reduce the measure to a single trait or characteristic. Failure to do this in the example citing the size of the college resulted in a misspecification of this variable. The respondents variously interpreted size of college

as spring enrollment, fall enrollment, total spring and fall enrollment, total full-time plus part-time spring enrollment, total full-time and part-time fall enrollment, full-time equivalent enrollment, and so on. The construct should have been operationalized as the enumeration of both the total full-time enrollment and the total part-time enrollment as of the close of the spring 1992 semester or quarter.

Developing Measures From Operationalized Constructs

Measures are developed by giving operationalized constructs a dimension. Measures qualify and sometimes quantify the trait in a single dimension such as presence or absence or the amount, intensity, value, frequency of occurrence, or the ranking or rating or some other form of comparative valuation or quantification. The next few paragraphs will help familiarize the reader with some of the requirements of a measure. Although this familiarization will proceed in other chapters of this paper through discussion and example, evaluators should consult a text specifically devoted to measurement or consult a specialist when complex measures are required.

Measures must be accurate, precise, valid, reliable, relevant, realistic, meaningful, comprehensive, and in some cases complementary, sensitive, and properly anchored. While evaluators may readily understand the meaning of precision and accuracy, some of the other terms may need to be defined, because in measurement they are used in a very special way. For instance, measures are considered valid if they are logical and they measure what they say they are measuring. They must adequately represent the trait in question. They must consistently predict outcomes, vary as expected in a variety of situations, and hold up against rigorous attempts to prove them invalid. We have all seen valid and questionable measures. Positive examples might be found in well-executed polls that predict voter outcome to a reasonable

degree of accuracy. A negative example might be found in the logic of using complaints as a measure of discrimination, because the cost, time to resolve a case, difficulty in proving discrimination, difficulty in filing, fear of retaliation, and other reasons discourage the aggrieved from filing a complaint.

Next, consider reliability, which is different from and independent of validity. To be reliable, a measure must give consistent results when repeated under similar situations. For example, IQ tests and employee attitude surveys usually give consistent results when repeated under similar circumstances with the same people.

Measures should be relevant, meaningful, and realistic. For example, some very valid measures like IQ and grade-point average are used to hire employees. These are not relevant measures if the new employee is expected to be creative and inventive and generate new ideas, because the traits of IQ, grade-point average, and creativity are not correlated. Also, the labels given to the measure should correctly describe and communicate its meaning. For example, managers frequently measure things like costs, staff time, and number of reports under the term "quality measures." These measures may index effectiveness or productivity but not quality. The measure should be realistic or practical. For example, if a reader's pupils are dilated, this might be a good measure of his or her interest, but these observations are very hard to make. Therefore, under certain circumstances, the accuracy of the respondents' information recall and self-reports, while not as accurate, are more useful because answers are easy to obtain.

Ideally, measures ought to be comprehensive and, in some cases, complementary. Comprehensive measures span the entire range of values that are of

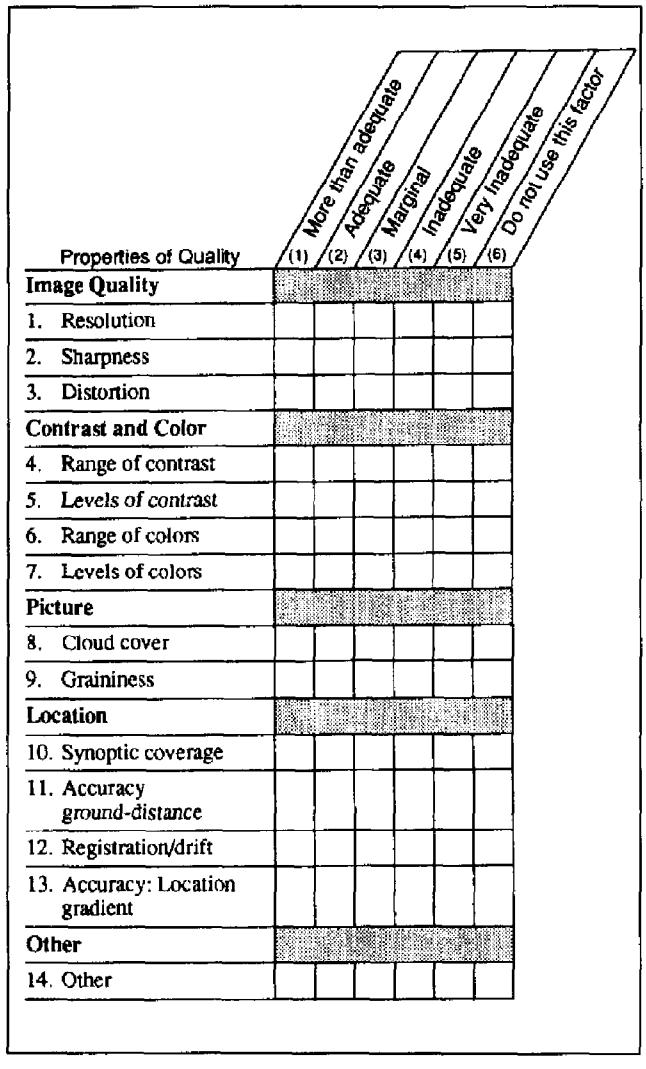
interest with equal precision. A single measure usually refers to a single trait, but sometimes if the construct is multidimensional or has several traits for reasons of economy or the need to capture two or more traits as they work together, we develop a measure that captures these multitrait effects. For example, asking the respondent if the text was easy to read and readily understandable might be considered a comprehensive measure. In contrast, complementary measures are measures that are distinct and must be taken together to reflect the construct. For example, the number of contrast shades and the sharpness of the contour lines are needed to measure photographic image quality.

Other features of measures that are also important are sensitivity and anchoring. Sensitivity refers to a measure's ability to detect (1) the presence or absence of the trait, (2) levels of intensity, or (3) changes in the level of intensity with sufficient precision at sufficiently low levels to meet the needs of the evaluation. Anchoring refers to the establishment of clear, concrete points on the measurement scale that are meaningful to the respondent. That is, the scale should have meaningful starting, interim, mid, and end points. For example, we might anchor estimations of lighting quality as too dim (not bright enough to read a newspaper), appropriate (could comfortably read a newspaper), or too bright (too much glare to comfortably read a newspaper).

An example of a complex measure taken from one of the cases cited in the preceding part of this chapter is presented in figure 2.1. The measure was developed from a construct identified with a questionnaire framework: the user's perception of the quality of an earth-orbiting satellite image. The construct was operationalized and developed into a measure of image quality. During this process, particular

attention was given to accuracy, precision, validity, reliability, realism of application, meaningfulness of concept, the comprehensiveness and complementary requirements, measure sensitivity, and anchoring of the measure.

Figure 2.1: Operationalized Variable in Question Response Format



Specify the Key Variable Relationships

We conclude this chapter with a brief but important discussion on specifying the variable relationships to be evaluated. (The two remaining sets of documentation needed to initiate the planning—the

identification of and the selection of the target population—are discussed in the next chapter.) This task is important because, as we shall see, errors or omissions in specifying the variable relationship can either invalidate or weaken the evaluation. In this task, evaluators document and review all variables to ensure that all key variable relationships are included and specified with common units of analysis and for appropriate functional relationships and in the appropriate measurement stratification and time periods so as to permit statistical, temporal, and cross-sectional observations and comparability. These variable relationships should be documented down to the level of measurement specification.

Then the evaluation design, the evaluation framework, and the questionnaire framework should be checked against this documentation to make sure nothing important is left out and that nothing unnecessary is included. A review should ensure that the sample or population measurements are to be taken on—and generalized to—common units of analysis. For example, in one case we found that one measure was to be taken on contractors, while its comparison measures applied to contracts. A review should be made for changes that would facilitate statistical comparability. For instance, the evaluators may find that one of the measures to be related is unnecessarily categorized while the other is continuous, or that some measures are inappropriately categorized for the intended cross sectional comparisons, thus weakening the statistical power of the analysis or, worse yet, rendering the analysis invalid.

Further, review should make sure the specified categories in the comparison variables are not likely to confound cross-sectional comparisons. For instance, suppose we know from past studies that the effects of training are not likely to be noticed until 9

months later, there is less bias against the mentally disabled in the city than in the suburbs, or treatment for violence exposure is most effective soon after the incident. If evaluators test for training effect soon after the training, they may not see an influence because the trainees did not have enough development time to assimilate their experience. If the test is for bias against the mentally disabled only in the inner city rather than in both the inner city and the suburbs, the evaluators may not find the effect because this bias is less noticeable in the inner city. If they test for the effects of treatment for exposure to violence on only those who waited a year before receiving treatment, they may not see the effect because the treatment was given too late to do much good. Hence, evaluators must make sure that the cross-sectional comparison categories are structured to capture, not hide, the effects under study.

Another point is to make sure the temporal comparisons are appropriate. For example, it is not unusual to find that the data for the different variables in the relationship are to be collected during different years. Finally, it is important to be sure important categories were not left out. This is because the sampling specialists will use this documentation to design the sample. For instance, in one case the evaluator was disappointed to find that the sample did not have enough power to compare important city, race, and educational stratifications because the sampling specialist had not been aware of these stratifications.

Chapter 3

Designing the Sample or Population for Data Collection

Along with deciding what to ask, evaluators must decide who to ask. The people questioned must have the information the evaluators seek, they must be readily identifiable and accessible, they must be willing and able to answer, and they must be representative of the population being measured. They can be migrant workers, prisoners, police, scientists, medical doctors, commanders or soldiers, inner city African American youths, or government officials.

Ideally, everyone in the population should be questioned, and sometimes this is done if the population is very small. But usually the best that can be done is to take a sample of these people and generalize the findings to the population they come from.

In theory, to generalize findings, evaluators must first define the population. Then they should enumerate every unit in the population in a way such that every unit has an equal chance of being selected for the sample. In practice, it may be unrealistic to expect to enumerate every unit in a real population (for example, all persons who participated in a government program such as Head Start), but the enumeration must be reasonably complete and accurate and be reasonably representative of the actual population. The evaluators must then draw a representative sample from this population.

Survey Population

However, the sample cannot be determined or drawn until the evaluators have studied the size and characteristics of the population they want to know about. All too often, this step in questionnaire development is overlooked or assumed to be routine. Then, when the questionnaire is complete and ready to be mailed, the team is faced with weeks of hard

research, or a major redesign, because the sample was not well founded.

The first step in defining the survey population is to learn about the population distribution—the major categories of units and the numbers in each category. For example, if the evaluators want to sample banks, they should learn the differences between county, regional, statewide, branch, and unit banks; they should know geographic location factors and understand the basis for classifying banks as very large, large, medium, and small. If they are studying unit commanders in the armed services, they should know the unit sizes and types and the variations among the services. This research will help in designing sampling factors, such as stratification and stratification size, and will ensure a representative sample.

Once the evaluators are familiar with the characteristics of the population, they can look for sources that enumerate each unit in the population or develop a reasonable theory for selecting the sampling units. The enumeration should be accurate, up-to-date, and organized to reflect the distribution characteristics. Sometimes this task is relatively easy. For example, in one project we needed to assess the effect that the Foreign Corrupt Practices Act had on U.S. business. The act prohibits payments to foreign officials if the purpose is to influence business. The population was U.S. companies that conduct most of the foreign business. These companies were readily identified because they were among the Fortune 1,000 companies, which conduct most of the foreign business. All we had to do was buy this list from Fortune magazine. The list gave the order of the companies by sales volume and provided information on each company's activities and the name and address of both the chief executive officer and the chairman of the board. However, for many other

projects, considerable effort is needed to document the survey population.

In practice, evaluators rarely have a list of the real population; at best they have only a list at the time the source material was current. By the time the questionnaire is administered, some units will have left the population and others will have joined it. For example, in the Fortune 1,000 evaluation, 6 percent of the firms left the population and we do not know how many may have joined it. The sample analysis must evaluate and make statistical adjustments for the losses. Whenever possible, the effect of the additions should also be considered.

The best way to start enumerating a population is to talk to experts in the field and search out likely organizations, archives, directories, libraries, and management information systems until a reliable source has been discovered. Then the sampling units or population elements are organized, reorganized, or indexed into groups or frames, so they can be reached by a random, systematic, or prescribed process. For example, in one evaluation, we had to locate retired military users of military medical facilities. From a Department of Defense archival data base we were able to get the names and addresses of all the retired military personnel but we had no way of knowing if they were users of a particular medical facility. Our field work showed that retired military were likely to travel up to 40 miles to use hospital services; if they lived farther away, they usually made other arrangements. So we developed a computer program, based on zip codes, that matched persons to the hospitals that were within 40 miles of their homes.

In a study of zoning problems encountered by group homes for the mentally disabled, we discovered that there was no national register of group homes. Since this was a study to see if this restrictive zoning

practice was geographically widespread, we sampled catchment areas. We then called up the catchment area directors and got the names and addresses of every group home in each catchment area and sent the group home directors a questionnaire asking about their zoning problems.

Sometimes, no matter how hard the search, archival data or records cannot be found from which to develop a population. When this happens, the best thing to do is to look for groups, sections, or clusters of files or lists that contain the information. Or the evaluators may want to look at existing data to surmise some ratio or relationship associated with the population. For example, if they want to define the population of general aviation flight-service airport specialists, they may be able to use previous work or pilot or survey studies. For example, from previous experience, they may find that they can estimate that the average number of specialists per airport is 16, multiply 16 specialists by the 316 airports, and estimate the population at about 5,000.

Unfortunately, in a great many cases, there is neither a population enumeration nor a way to get cluster, unit, or ratio figures. In these cases, the evaluators must try to document the biggest possible portion of the most important and most representative cases, or they must develop some reasonable theory for selecting the sampling units. For example, to get a representative list of internal auditors, the evaluators might use the membership list for the Institute of Internal Auditors plus a list of the internal audit departments for the Fortune 1,000 companies. The latter would be included because most of them have internal audit departments.

In one situation, we had to sample major importers and exporters. The available list had over 10,000 entries, almost all of which were too small to be

considered major. So we used a combination of a "small world network" and a "snowball" approach. We found an association on the eastern coast to which most major mid-Atlantic shippers belonged. We contacted the association and obtained a list of the major shippers and their business volume. This association identified two other shippers' associations, which provided their lists and the names of six more associations. We continued until we had identified all associations and had a list of most of the major shippers. The shippers' associations reviewed our list and estimated that it accounted for 82 percent of the import-export business.

Many other sources of specialized lists are available, but their reliability varies considerably. For example, major organizations such as the American Medical Association, the National Education Association, and the National Association for Home Builders can provide detailed address lists and population descriptions of their members. However, their cooperation varies with their interest in what the job is about. The cost for lists can be anything from nothing to a few hundred to several thousand dollars. Although the Bureau of the Census sometimes has useful lists, such as the census of manufacturers and the census of governments, these sources may be out of date. Many commercial sources, such as Ruben and Donnelly, Polk, and Thomas, sell population lists. Also, some commercial firms such as Dunn and Bradstreet sell specialized lists for various users, such as mail order companies. Care must be taken in using these lists because their quality varies considerably and very little may be known about the bias built into them, how they were developed, or what they include and, more importantly, exclude.

Before using a list, it is a good idea to review and perhaps test it. For example, in a sample survey of farmers, the address list was developed from a list of

subscribers to the Farm Home Journal. The list turned out to be several years old, and many of the subscribers were not farmers in the technical sense but people who sold or bought agricultural equipment or products or who were interested in rural living.

Selecting the Sample

Once the population has been enumerated and the evaluators are sure that it represents the population to which they want to generalize, they are ready to draw the sample.

The sample must be drawn in accordance with a procedure that ensures a random selection. The sample size must be large enough to provide the degree of measurement precision and accuracy generally accepted by the scientific community. This must be done very efficiently and cost effectively. In many instances, accomplishing this will require the assistance of a sampling statistician who has the appropriate technical skills and practical experience.[1]

Nonstatistical Sampling

Questionnaires may be used on projects in which statistical sampling is not used, so we need to consider briefly other ways in which evaluators select cases (Deming, 1960). Either all the cases can be studied—that is, a census can be taken—or part of the population can be selected in a nonstatistical manner. When evaluators take part of the population, they usually do so for a reason. It may be that they are doing a case study, so they select one or more cases that provide the best opportunity to observe the phenomena or relationships of interest, and they do not need to generalize their findings to the population. In other situations, the evaluators know

[1]See U.S. General Accounting Office, Using Statistical Sampling, GAO/PEMD-10.1.6 (Washington, D.C.: May 1992). This paper provides a thorough treatment of this topic.

very little about the population and cannot draw a statistical sample, so they arbitrarily select as many cases as they can and report the findings. However, in many situations, evaluators want to generalize and they know something about the population but it is just not feasible to draw statistical samples. So they pick a sample that they hope will correspond, in its features, to the population, even though they know they will not be able to use the powerful reasoning associated with statistical samples. An important category of nonstatistical sampling is "judgment sampling."

A judgment sample draws its name from the fact that in the judgment of the evaluator, the cases chosen correspond to certain aspects of the population. The cases may be selected because they are judged most typical, because they represent the extreme ranges, because they represent a known part of the population, or because they simulate or act as a proxy for a representative sample from the population. For example, we could interview all the Fortune 500 chief executive officers in New York and Chicago because we believe that this sample is typical of chief executive officers in large companies. We could study selected group homes for the mentally disabled in California, Mississippi, New York, and Texas, because these states represent the extremes of the laws and practices. We could study 50 prime contractors with the Department of Defense in California and New York, because these contractors account for 82 percent of all defense contracts. We might pick 15 airports in 11 states, such that the sample would be similar to the population of airports with respect to size, geographic coverage, and weather conditions.

As a rule, the use of judgment sampling in a project in which the intent is to generalize is ill advised, because arguments to support generalization cannot be nearly as persuasive as with statistical samples. However,

occasions may arise (as with a very homogeneous population) in which the situation is not altogether bleak.

When the validity of the findings depends on the extent to which they can be generalized to the population, and when there is no statistical sample, it might help to have some rule of thumb that might compare judgment samples to statistical samples. One way to picture the relationship between statistical samples and judgment samples with respect to representativeness might be to imagine a credibility scale from 1 to 10. Assume that a score of 1 is the value given to a single case study designed without intent whatsoever to generalize, and 10 is the credibility associated with studying the whole population. A very large, statistically valid random sample might yield a value of 9. A large, medium, and very small but statistically valid random sample might yield respective scores of 8, 7, and 6. If we made many case studies but did not take a random sample, we might get a value of 4. We might extend this value to 5 if the groups were large enough to provide statistical certainty within their limited area of selection or if the population was very homogeneous. We might get the same score of 5 if we selected a number of cases that represented the range of conditions and circumstances that apply to the population. (Incidentally, this is how pretest candidates are selected, because there is neither time nor resources to draw a statistically valid sample.) However, the score would drop to 3 or even 2 if we selected many or fewer cases without giving consideration to representing the expected range of conditions.

A few years ago, we did a review of the elderly in which we selected thousands of cases at random from the same city. This might have been acceptable, from a generalization viewpoint, if we were measuring the conditions associated with cholesterol levels; these

levels could be presumed similar for most U.S. city-dwellers. However, in this review, we were concerned about programs and their effects, which may have varied from city to city. Thus, limiting the sample to one city prohibited generalizations beyond the city that was studied. Another example involved a population of 132 health maintenance organizations. We arbitrarily picked 16 of these organizations and collected data from hundreds of people in each one. In the end, what we came up with was a set of 16 case studies. Although the sample for each case study was representative of the population of people in one of the 132 health maintenance organizations, the 16 case studies together permitted only very careful and limited findings. We might have had a much more powerful evaluation at a fraction of the cost if we had taken a random sample of organizations and looked at fewer cases within each organization.

Chapter 4

Formatting the Questions

Before writing the questionnaire, the evaluators need to choose the format for each question. Each format presented in this chapter serves a specific purpose that should coincide with the available information and data analysis needs.

Open-Ended Questions

Open-ended questions are easy to write and require very little knowledge of the subject. All the evaluators have to do is ask a question, such as "What factors do you consider when you pick a carrier?" But this type of question provides a very unstandardized, often incomplete, and ambiguous answer, and it is very difficult to use such answers in a quantitative analysis. Respondents will write some salient factors that they happen to think of (for example, lower rates and faster transit time) but will leave out some important factors because at that moment they did not think of them. Open-ended questions do not help respondents consider a range of factors; rather, they depend on the respondents' unaided recall. There is no way of knowing what was important but not recalled, and because not all respondents consider the same set of factors, it may be extremely difficult or impossible to aggregate the responses.

Also, the evaluators may not know how to interpret the answers. For example, people might say they choose a carrier because it is more convenient or less trouble. There is no way of knowing what this means. It may mean any thing from faster transit time to easier documentation.

Another problem is that open-ended questions cannot easily be tabulated. Rather, a complicated process called "content analysis" must be used, in which someone reads and rereads a substantial number of the written responses, identifies the major categories of themes, and develops rules for assigning responses to these categories. Then the entire sample has to be

gone through to categorize each answer. Because people interpret differently, three or four people have to categorize the answers independently. Furthermore, rules must be developed to handle disagreements and only very low levels of qualitative analysis can be performed.[1] Similarly, at the conclusion of the data reduction phase, only very low levels of qualitative analysis can be performed.

Still another problem is that open-ended questions substantially increase response burden. They usually take several minutes to answer, rather than a few seconds. Because respondents must compose and organize their thoughts and then try to express them in concise English, they are much less likely to answer.

However, open-ended questions do sometimes have advantages. It may happen that they are unavoidable when, for example, we are uncertain about criteria or we are engaged in exploratory work. If we ask enough people an open-ended question, we can develop a list of alternatives for closed-ended questions. We can also use open-ended questions to make sure our list of structured alternatives did not omit an important item or qualification. We can also ask open-ended questions to obtain responses that might further clarify the meaning of answers to closed-ended questions or to gather respondent examples that can be used to illustrate points. The rest of this chapter details closed-ended questions, because they are the meat and potatoes of our work.

Fill-in-the-Blank Questions

Each questionnaire usually has some fill-in-the-blank questions. They are not open-ended because the blanks are accompanied by parenthetical directions

[1]Interrater reliability is a measure of the consistency among the people categorizing the answers.

that specify the units in which the respondent is to answer. Some examples are shown in figure 4.1.

Figure 4.1: Fill-in-the-Blank Questions

1. What was your age on your last birthday?_____(age in years)

2. What was your city's infant mortality rate last year? _____(mortality rate in deaths/1,000)

3. What size is your printing plant?_____(in square ft.)

Fill-in-the-blank questions should be reserved for very specific requests. The instructions should be explicit and should specify the answer units. Sometimes, several fill-in-the-blank questions are asked at once in a row, column, or matrix format, as shown in the examples presented in figure 4.2.

Figure 4.2: Fill-in-the-Blank Row, Column, and Matrix Formats

1. Estimate the number of children, juveniles, or adults that you usually care for at any one time. *(Answer for each appropriate age group.)*

Age Group	Number of children, juveniles, or adults
1. Under 6 years of age	
2. From 6 to under 9	
3. From 9 to under 12	
4. From 12 to under 15	
5. From 15 to under 18	
6. From 18 to under 21	
7. Adults over 21 years	

2. For the countries listed above, identify the countries or territories in which you are doing exploratory work and describe the extent of the exploratory activities—that is, the size of the area under exploration, line miles of seismic data, size of area under drilling rights and number of exploratory wells.

	Exploration activities				
	Countries doing exploration	Square miles under exploration	Line miles of seismic data	Square miles under drilling rights	Number of exploratory wells
1.					
2.					
3.					
4.					
5.					
6.					

3. Indicate the dollar amount of nonfederal funding (city, county and state) provided to your project for each of the following six grant years.

Dollar amount of nonfederal funds in project by source (Specify the dollar amount in thousands.)

	City	County	State
1. 1st grant year			
2. 2nd grant year			
3. 3rd grant year			
4. 4th grant year			
5. 5th grant year			
6. 6th grant year			

Yes-No Questions

Unfortunately, yes-no questions are very popular. Although they have some advantages, they have many problems and few uses. Yes-no questions are ideal for dichotomous variables, such as black and white, because they measure whether the condition or trait is present or absent. They are therefore very good for filters in the line of questioning and can be used to

move respondents to the questions that apply to them, as in figure 4.3.

Figure 4.3: Yes-No Filter Question

1. Did you get training? *(Check one.)*

   1. ☐ Yes (continue)
   2. ☐ No (go to question 5)

However, most of the questions GAO asks deal with measures that are not absolute or measures that span a range of values and conditions. Consider the question: "Were the terms of the contracts clear?" Most people would have trouble with this question because it involves several different considerations. First, some contracts may have been clear and others may not have been. Second, some contracts may have been neither clear nor unclear or of marginal clarity. Third, parts of some contracts may have been clear and others not clear.

Because so little information is obtained from each yes-no question, several rounds of questions individually have to be administered to get the information needed. "Did you have a plan?" "Was the plan in writing?" "Was it a formal plan?" "Was it approved?" This method of inquiry is usually so boring as to discourage respondents.

Sometimes, question writers try to compress their line of inquiry and cause serious item-construction flaws. They ask for two things at once—a double-barreled question. For instance, a yes-no answer to "Did you get mission and site support training?" is imprecise.

How do respondents answer if they got mission but not site support training?

A related question-writing mistake is mixing yes-no and multiple choice. See figure 4.4.

Figure 4.4: Mixed Yes-No and Multiple Choice Question

1. Did you get mission and/or site training?
   1. ☐ Yes, mission but not site training
   2. ☐ Yes, site but not mission training
   3. ☐ Yes, both mission and site training
   4. ☐ No, neither mission nor site training

The example in figure 4.4 has several problems. The question and the response space do not agree. This slows up the cognitive processing because the question prepares the reader for a simple yes-no answer. But in reality the reader gets not a yes-no answer space but, rather, a list of qualified alternatives. The response alternatives are biased toward "yes" because most of the choices have "yes" in them. Furthermore, "no" in the last item cannot be used with the correlative conjunction "neither nor," because this is an unintended double negative. Such questions make a simple inquiry difficult because they are counter to the cognitive process, burdensome, and cause errors.

Yes-no questions are prone to bias and misinterpretation for several reasons. First, many

people like to say "yes." Some have the opposite bias and like to say "no." Second, questions such as "Do you submit reports?" have what is called an "inferred bias" toward the "yes" response. The most common way to counter this bias is to add the negative alternative—for example, "Do you submit reports or not?" However, if this is done, the use of yes-no choices in the answer must be qualified or avoided. Without this precaution, a simple "yes" answer may be read as applying to both parts of the question, "Yes, I submit" and "Yes, I do not submit." A simple "No" might also be read as "No, I do not submit"—a double negative. To prevent confusion, qualify the answer choices or avoid yes-no answers. See figure 4.5.

Figure 4.5: Balanced and Unambiguous Yes-No Question

1. Do you submit reports or not? *(Check one.)*

   1. ☐ I submit reports.
   2. ☐ I do not submit reports.

"Implied No" Choices

In figure 4.6, failure to check an item implies "no." The implied-no choice format is used because it is easy to read and quick to answer.

Figure 4.6: "Implied No" Question

1. What health problems, if any, did the VA tell you that you had? *(Check all that apply.)*

   1. ☐ Skin problems
   2. ☐ Liver or kidney problems
   3. ☐ Tumors or growths
   4. ☐ Problems with your nerves
   5. ☐ Other health problems (please specify)

When evaluators want to emphasize the "no" alternative, they can expand the implied-no format to include one column for "yes" answers and one for "no." "No" is listed as an option when the respondent might not answer or might overlook part of the question, as when the choices are difficult, the list of items is long, or the respondent's recollection is taxed. If "no" is not included as an alternative, no's will be overreported, because the analysts will not be able to differentiate real no's from omissions and nonresponses. An example appears in figure 4.7.

Figure 4.7: Emphasized-No Question

1. Did the VA ask if you had the following health problems during or since your service in Vietnam or not? *(Check one column for each row.)*

Questions asked	Yes (1)	No (2)
1. Nervousness		
2. Headaches		
3. Numbness in arms, legs, hands, feet		
4. Infections		
5. Liver problems		
6. Weight loss		
7. Fatigue		
8. Skin problems		
9. Lung problems		
10. Change in sex drive		
11. Sterility		
12. Birth defects in children		

Single-Item Choices

In single-item choices, respondents choose not "yes" or "no" but one of two or more alternatives. See figure 4.8 for an example. Since yes-no and single-item choices are similar, they have the same types of problems, but the difficulties are less pronounced in some respects and accentuated in others.

Figure 4.8: Single-Item Choice Question

1. There are two programs for educating the handicapped. One program provides special education in separate classrooms and uses a curriculum different from that used for the main group of children. Another program (called mainstreaming) includes the handicapped in the regular classroom, adapts the main curriculum to special education, and makes other provisions for the handicapped. The question is, which alternative do you prefer? *(Check one.)*

   1. ☐ Separate special education classes
   2. ☐ Mainstream classes

On the positive side, the differences between the choices are usually clear, and the writer can set up a truly dichotomous question. If used carefully, the single-item choice can be efficient. It often serves to filter people out or to skip them through parts of the questionnaire. It is not likely to be overused and cause excessive cycles of repetition. Furthermore, the question writer is not likely to compress the question into a double-barreled item. The single-choice format is also not subject to bias from yea sayers or nay

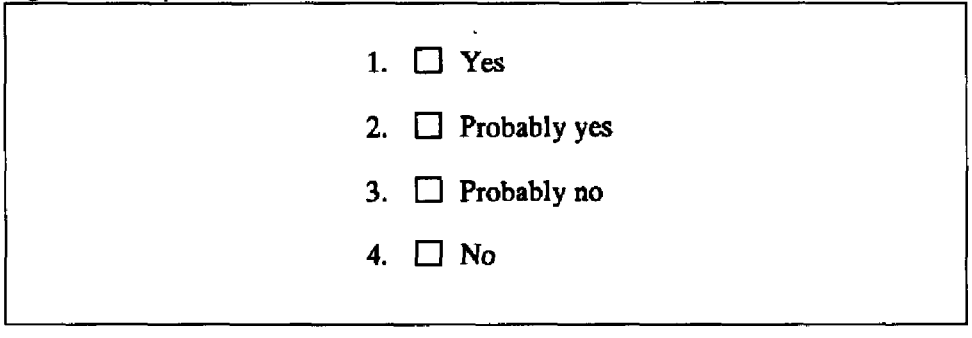
sayers. And eliminating the negative alternative reduces misinterpretation.

But there are problems. In the single-choice format, the writer is more apt to bias one of the choices by understating or overstating it. Some writers may not properly emphasize the second alternative; others, aware of this tendency, overcompensate.

Expanded Yes-No Questions

One way around the yes-no constraints is to use an expanded yes-no format like that shown in figure 4.9. The expanded yes-no format gives a measure of intensity, avoids some of the biases common to yes-no, implied-no, and single-choice questions, and resolves the problem of quibbling. Consider the question, "Could you have gotten through college without a loan or not?" Also in the expanded format more students will answer in the negative than otherwise.

Figure 4.9: Expanded Yes-No Format



The expanded alternatives can have qualifiers other than "probably yes" and "probably no." Qualifiers can be changed to meet the situation—"generally yes" and

"generally no" or "for the most part yes" and "for the most part no."

Free Choices

Yes-no, implied-no, single-choice, and expanded formats are forced choices in that respondents must answer one way or the other. Forced-choice items generally simplify measurement and analysis because they divide the population clearly into those who do and those who do not or those who have and those who have not. Unfortunately, putting the population into just two camps may also oversimplify the picture and yield error, bias, and unreliable answers. To avoid this problem and to reduce the respondent's burden, a middle category can be added, as in the question in figure 4.10.

Figure 4.10: Expanded Yes-No Format With Middle Category

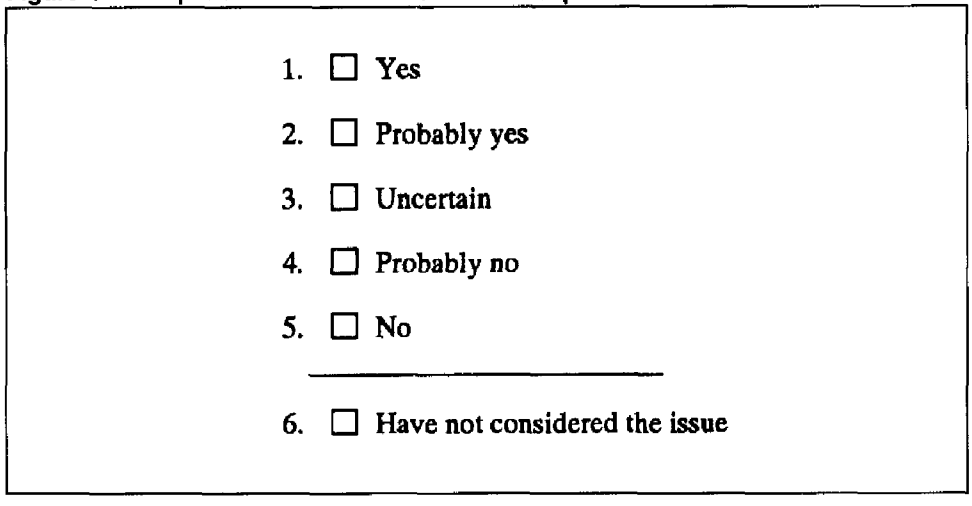
1. ☐ Yes
2. ☐ Probably yes
3. ☐ Uncertain
4. ☐ Probably no
5. ☐ No

Even though the proportion of yes's to no's will not change, the evaluators will have a better measure of the yes-no polarization, because the middle category absorbs those who are uncertain. A good rule of thumb is that if we are not certain that nearly

everyone can make a clear choice—we include a middle category.

Usually, the question asker will also put in an "escape choice" to filter out those for whom the question is not relevant. Examples are "not applicable," "no basis to judge," "have not considered the issue," and "can't recall." See figure 4.11.

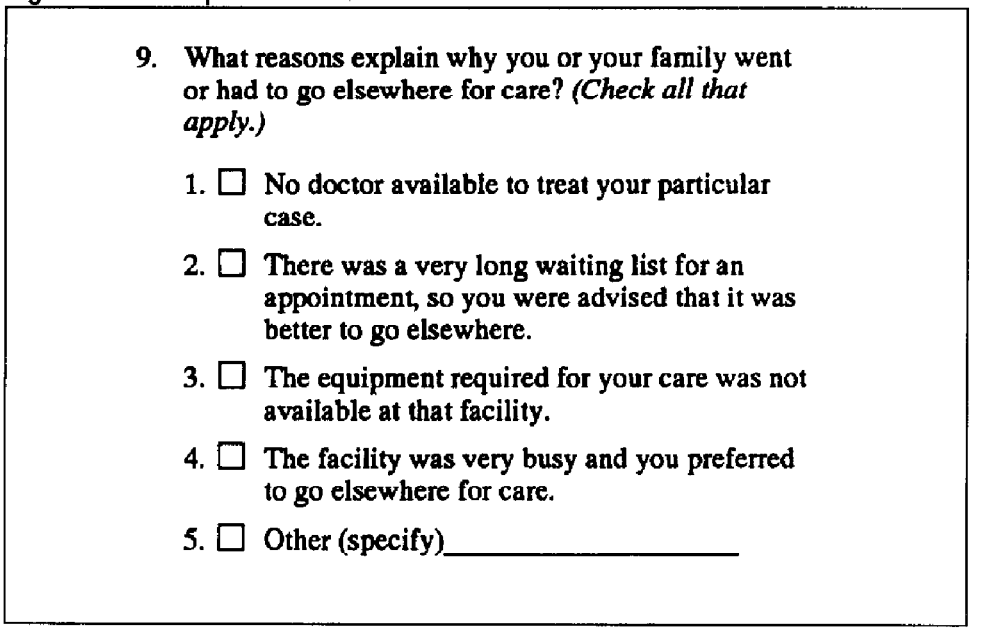
Figure 4.11: Expanded Yes-No Format With Escape Choice



Multiple-Choice Questions

The most efficient format—and the most difficult to design—is the multiple-choice question. The respondent is exposed to a range of choices and must pick one or more, as in the example in figure 4.12.

Figure 4.12: Multiple-Choice Question



Multiple-choice questions are difficult to write because the writer must provide a comprehensive range of nonoverlapping choices. They must be a logical and reasonable grouping of the types of experience the respondents are likely to have encountered.

The example in figure 4.12 turned out to be flawed in practice. We learned during the pretest that we had left out some important choices. We detected this error because many respondents wrote answers in the "other" category.

Because this format is very important and requires the most research, field work, and testing, and because the analysis and interpretation can be

complex, we discuss multiple-choice question design in chapter 7 in considerably more detail.

Ranking and Rating Questions

Ranking questions are used to make very difficult distinctions between things that are of nearly equal value. The question forces the respondent to value one alternative over another no matter how close they are. The value that is assigned is a relative value. Rating questions are used when the alternatives are likely to vary somewhat in value and when evaluators want to know how valuable the alternative is rather than if it is a little more or less valuable than the next alternative. First consider ranking. In ranking, the respondents are asked to tell which alternative has the highest value, which has the second highest, and so on. They rank the choices with respect to one another, but their answers tell little about the intrinsic value of their choices. For example, suppose we asked respondents to rank the importance of the following services for institutionalized children: education, health care, lawn care, telephones, and choir practice. They would be hard put to choose between education and health care, because both are essential to the children's development. But they would have to rank one first and one second. Telephones would probably be ranked third. Compared to health care and education, telephones are much less important, yet they are ranked third just behind two services that are so important that it is difficult to choose between them.

Ranking starts to get hard for people when there are more than seven categories. This is because they can usually pick the first and second and third and then the last and next to the last and the next to the next to last, so that what is left is the middle. But for more than seven items, respondents begin to lose track of where they are with respect to the first, last, and middle positions. When this happens, they make

mistakes. For more than seven items, respondents can be given special task-taking procedures to counter this problem. But this procedure is rather burdensome.

Also, ranking questions have to be written very carefully. The slightest lapse in clarity in the question or the instruction given will cause some people to rank in the reverse order or to assign two alternatives the same rank or to forget to rank every alternative. Nonetheless, ranking must sometimes be used. The example in figure 4.13 is one that has worked reasonably well. Respondents will make a few errors, but statistical procedures are available to handle them.

Figure 4.13: Ranking Question

Consider each of the following types of findings, which are often used to assess programs. FROM YOUR EXPERIENCE, which types of results do you think are more likely to impress the state education agency program (SEA) officers? Indicate your answer by rank ordering each of the following alternatives from the most to the least impressive. Select the type of result you think is most likely to impress the SEA officials. Rank this 1st by checking. Do the same for all the remaining categories, ranking them 2nd, 3rd, 4th, 5th, 6th, and 7th.

	1st	2nd	3rd	4th	5th	6th	7th
1. Improvement in educational management or accountability							
2. Improvement in school or facilities							
3. Student improvement *through gain scores* on grades or teacher rating							
4. Student improvement through gain scores on standardized norm referenced tests							
5. Student improvement through gain scores on criterion referenced tests							
6. Student improvement through gains in the affective domain (e.g., likes, dislikes)							
7. Improvement in curriculum and instruction							

Rating questions are perhaps our most useful format because we usually want to know the actual or absolute value of the trait we are measuring. Ratings are assigned solely on the basis of the score's absolute position within a range of possible values. For example, a rating scale might be assigned the following categories: of little importance, somewhat important, moderately important, and so on. In writing rating questions, we should try to categorize the scales in equal intervals and anchor the scale positions whenever possible. Aside from the scaling, rating questions are easier to write properly and cause less error than ranking questions. We can see from the two examples of the rating format shown in figure 4.14 that ratings provide an adequate level of quantification for most purposes. We can also see by comparing the examples in figures 4.13 and 4.14 that rating formats are far less cumbersome than ranking formats.

Figure 4.14: Rating Questions

1. Under what risk classification should Presentence Investigation reports contain recommendations for court conditions? *(Check one.)*

   1. ☐ Maximum risk
   2. ☐ Moderate risk
   3. ☐ Minimum risk

2. Rate how well the report contents were supported by verification, referencing of sources, statistics, statements of scientific certainty, or soundness of data-gathering methods. *(Check one.)*

   1. ☐ More than adequate
   2. ☐ Generally adequate
   3. ☐ Of marginal or borderline adequacy
   4. ☐ Inadequate
   5. ☐ Very inadequate

Guttman Format

In questions written in the Guttman format, the alternatives increase in comprehensiveness; that is, the higher-valued alternatives include the lower-valued alternatives.

Applying this principle in one job, we asked state resource officials how they benefited from an earth-orbiting satellite. The question is given in figure 4.15. Here we assumed that if respondents had measured the benefit, they had identified it, and if they had determined the cost-benefit ratio, they had

measured the primary and secondary benefits and lack of benefits as well as the worth or dollar value of these benefits and lack of benefits.

Figure 4.15: Guttman Question

Consider the benefits, if any, your state government may have received from participating in the LANDSAT program. Identify the benefit areas and the degree to which you can qualify and/or quantify these benefits. (Check column 1 if particular benefit not identified; otherwise check one of the columns 2-5.)

	Qualification of Benefits				
Benefit area	No benefit identified (1)	Identified benefits (2)	Measured some or all benefits (3)	Assessed worth and/or dollar value of benefits (4)	Made cost-benefit analysis (5)
1. Agriculture/forestry, range resources					
2. Land use survey and mapping					
3. Mineral resources, geostructural, and land form surveys					
4. Water resources					
5. Marine resources and ocean surveys					
6. Meteorology					
7. Environment					
8. Other					

Intensity Scale Questions

The intensity scale format is usually used to measure the strength of an attitude or an opinion. Two popular versions, the extent and expanded yes-no scales, are presented in figures 4.16.

Figure 4.16: Extent Scale and the Expanded Yes-No Scale Questions

1. To what extent, if at all, do you believe an international agreement against bribery would strengthen American companies' competitive position abroad? *(Check one.)*

   1. ☐ To little or no extent
   2. ☐ To some extent
   3. ☐ To a moderate extent
   4. ☐ To a great extent
   5. ☐ To a very great extent
   6. ☐ No opinion

2. Do you feel that an international trade agreement against bribery would strengthen American companies' competitive position abroad or not? *(Check one.)*

   1. ☐ Yes
   2. ☐ Probably yes
   3. ☐ Uncertain
   4. ☐ Probably no
   5. ☐ No

Likert Scale

Another frequently used intensity scale format is the Likert or agree-or-disagree scale. The Likert scale is easy to construct. Consider the extent-scale example of figure 4.16. As shown in figure 4.17, all the question

writer has to do is convert the question into a statement and follow it with agree-or-disagree choices.

Figure 4.17: Extent Scale Converted to Likert Scale Question

1. **An international agreement against bribery would strengthen U.S. companies' competitive position abroad.** ***(Check one.)***

   1. ☐ Strongly agree
   2. ☐ Agree
   3. ☐ Undecided
   4. ☐ Disagree
   5. ☐ Strongly disagree
   6. ☐ No basis for judging

However, if the writer is not careful, the simplicity and adaptability of the Likert scale format are often paid for by greater error and threats to validity.

First, there is bias. The Likert scale presents only one side of an argument, and some people have a natural tendency to agree with the "status quo" or the argument presented. Writers of Likert scale questions could attempt to counter this bias error by presenting the converse statement also. For example, they would first ask for a response to "My boss does not let me participate in decisions (agree or disagree)." Then in a

subsequent part of the questionnaire, they have to ask their questions in reverse: "My boss lets me participate in decisions (agree or disagree)."

But now the line of inquiry is no longer concise or simple. The questions are doubled in number with a serial repetitive format that interferes with the cognitive recall process, aside from inhibiting motivation because these formats quickly become boring. Furthermore, developing precise converse statements of counterbalancing intensity can be difficult and complex. For example, "not satisfied" is not necessarily the opposite of "satisfied." And in the example above, the phrase "My boss does not let me participate" is much more negative than the phrase "My boss lets me participate" is positive.

Another problem is that the extent of the respondent's agreement or disagreement with a statement may not correspond directly to the strength of the respondent's attitude about the Likert statement posed in the question. The respondent may consider the statement either true or false and respond as if the question were in an "either or" format rather than a graduated scale measuring the intensity of a belief.

The Likert question uses the statement as a reference point or anchor. Hence, what is measured may be not the strength of the respondent's attitude over the complete range of intensities but, rather, the range of intensities bounded or referenced by the position of the anchoring statement at one end of the range and unbounded at the other end of the range. To complicate things even more, the single-bounding anchor may not be at the extreme end of the range; this makes comparisons among items very difficult.

The point is that the indirect approach in the Likert scale may produce misleading results for a variety of

reasons. It is usually better to use a direct approach that measures the strength of the respondent's actual attitude over a complete range of intensities. For example, it is better to reformulate the item from "My boss never lets me participate" to "To what extent, if at all, do you participate?"

However, one situation in which the Likert scale is very useful is when extent of agreement or disagreement is closely and directly related to the statement. For instance, the respondent may be asked about the extent to which he or she agrees or disagrees with a policy, as in figure 4.18.

Figure 4.18: Likert Question Used to Evaluate Policy

1. Some people agree with GAO's policy on rotation, while others do not. The question is, how do you feel about the policy? *(Check one.)*

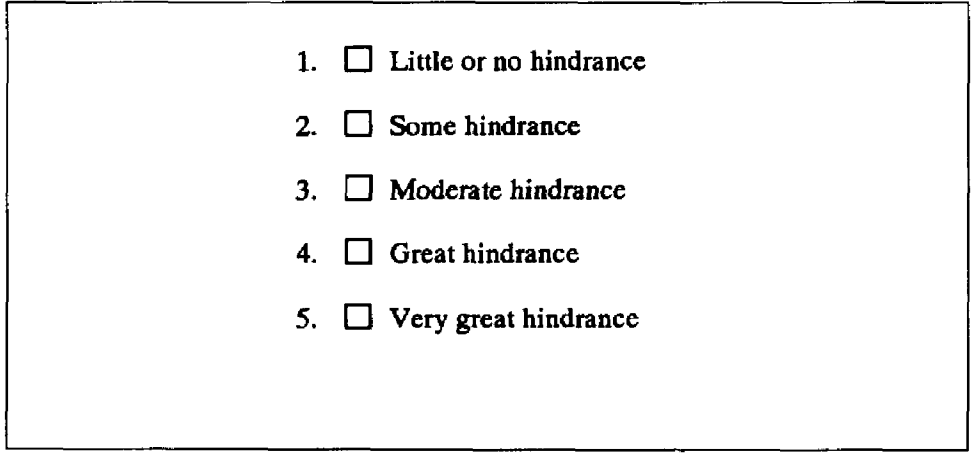
   1. ☐ Strongly agree
   2. ☐ Agree **more** than disagree
   3. ☐ Undecided
   4. ☐ Disagree **more** than agree
   5. ☐ Strongly disagree

Amount and Frequency Intensity Scales

Many questions ask the respondent to "quantify" either amounts or frequencies. These are relatively simple. They use certain descriptive words to characterize the amount, frequency, or number of items being measured. For example, traits like "help," "hindrance," "effect," "increase," or "decrease" can be quantified by adding "little," "some," "moderate,"

"great," or "very great." Certain adjectives like some and great have a stable and relatively precise level of quantification. For instance some is usually considered to be about 25 percent of the amount shown on the scale and a great amount is usually considered to be about 75 percent. Sometimes such adverbs as "very" and "extremely" are used. Quantities can also be implied by the sequence of numbered alternatives ordered with respect to increasing or decreasing intensity. See figure 4.19, which uses both methods together, in the common practice.

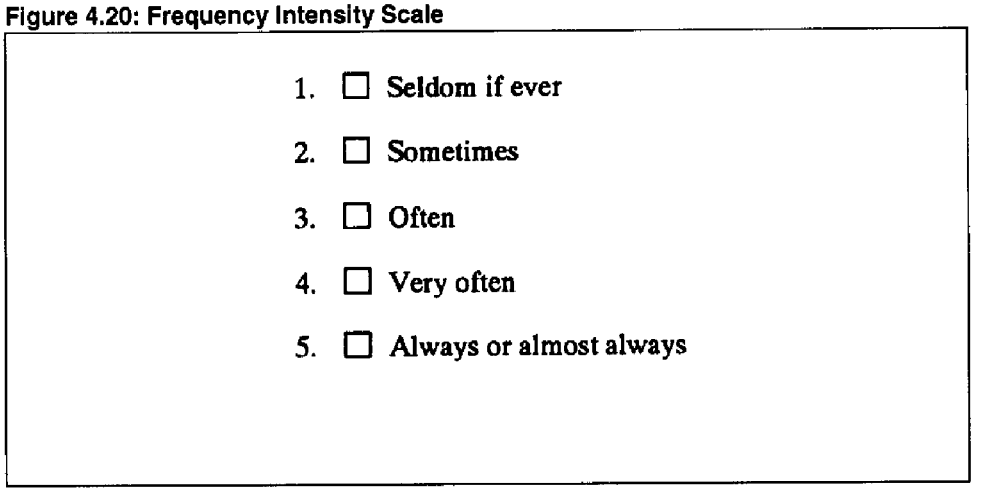
Figure 4.19: Amount Intensity Scale



Frequencies or occurrences of events are treated the same way. Question writers know that words like "sometimes" and "great many" or "very often" mean about one fourth of the amount or 25 percent of the time and three fourths or 75 percent of the time, respectively, to most people. Similarly, words like "about half" and "moderate" anchor the midpoints. As

with amount intensity scales, it is important to use both numbered, ordered scalar presentations and words to quantify the scale intervals. See figure 4.20.

Figure 4.20: Frequency Intensity Scale



In many amount and frequency measures, where ambiguities are likely to occur, it is also important to use proportional anchors such as fractions and percents or verbal descriptive anchors such as once a day or once a month in addition to the adjective and scale number anchors. Examples are shown in figure 4.21.

Figure 4.21: Frequency and Amount Intensity Scales With Proportional and Verbal Descriptive Anchors in Addition to the Conventional Adjective and Scale Number Anchors

1. ☐ Seldom if ever (0 to 10% of the time)
2. ☐ Sometimes (about 1/4 of the time)
3. ☐ Often (about 1/2 of the time)
4. ☐ Very often (about 3/4 of the time)
5. ☐ Always or almost always (90 to 100% of the time)

2.

1. ☐ Always or almost always (once a day or so)
2. ☐ Very often (every other day or so)
3. ☐ Often (about once a week)
4. ☐ Sometimes (every two or 3 weeks)
5. ☐ Infrequently (once a month or less)

3.

1. ☐ To little or no extent; less than 10% of the streams are covered
2. ☐ To some extent; perhaps 1/4 of the streams are covered
3. ☐ To a moderate extent; about half the streams are covered
4. ☐ To a great extent; about 3/4 of the streams are covered
5. ☐ To a very great extent; nearly all of the streams are covered

Branching Intensity Scale Formats

So far, all the examples have illustrated nonbranching formats. However, even more precise measures can be obtained with branching formats. An example is shown in figure 4.22.

Figure 4.22: Branching Intensity Scale Format

1. If a group home for mentally ill were applying for a license in your neighborhood, would you support or oppose this licensing or are you undecided? *(Check one.)*

   1. ☐ Support (continue)
   2. ☐ Undecided (go to 4)
   3. ☐ Oppose (go to 3)

2. If you would support this licensing, to what extent would you support it? *(Check one.)*

   1. ☐ To a little extent
   2. ☐ To some extent
   3. ☐ To a moderate extent
   4. ☐ To a great extent
   5. ☐ To a very great extent

Fill-in-the-Blank Frequency Formats

Sometimes when evaluators have to be really precise and the range of frequency choices is very wide, such as in the study of repetitive behaviors, they can use a fill-in-the-blank format. What is asked for is the number of occurrences in a given time period or the

interval between events to be counted. Examples are shown in figure 4.23.

Figure 4.23: Number-of-Occurrences and Time Interval Formats

1. How many meetings have you attended in the last two full weeks?

_____(number of meetings in last two weeks, counting back from last full week)

2.
   1. When was the last meeting you attended?

   _____

   2. Before this meeting, how long had it been since you had attended a meeting?

   _____(number of days since attending another meeting)

Here are some guidelines for using intensity scales.

1. Pick a dimension and a dimension reference point; then decide whether the scale should increase in a negative direction from that reference point, increase in a positive direction, or both. For instance, consider the question, "To what extent, if at all, did the law

affect your business?" Here, the scale might go from reference point "no effect" to "a severe hardship" or, if you believe the law can only help, from "no effect" to "a very great help." But if the law could help some and hinder others, the scale would span the range from "a severe hardship" through the "no effect" reference point to "a very great help."

2. Use an odd-number of categories, preferably five or seven.

3. If there is a possibility of bias from the category ordering, order the scale in a way that favors the hypothesis you want to disconfirm and that disadvantages the hypothesis you want to confirm. This way, you confirm the hypothesis with the bias against you.

4. If there is no bias, start the scale with the most undesirable or negative effect and end the scale with the most positive categories.

5. Present the scale categories in the sequence that people are used to seeing them.

6. Pick scale-range anchors or poles (that is, specify the ends of the range) with concrete and unambiguous measures.

7. Use the item sequence and numbering to help define the range of categories.

8. Use words that are natural anchors or that will divide the scale at equal intervals, particularly over the middle two thirds or three fourths of the scale. For example, to most people, "some or somewhat" is usually perceived as about one fourth of the time, intensity, or amount, whereas "great" has a face value of about three fourths.

9. Anchor the intervals with numbers, fractions, or proportions and descriptions, when feasible.

10. Use a branching format when feasible, as it is precise.

Semantic Differential Intensity Scales

In a semantic differential question, frequencies or values that span the range of possible choices are not completely identified; only the extreme value or frequency categories are labeled. An example is shown in figure 4.24. The respondent must infer that the range is divided into equal intervals. The range seems to work much better with seven categories than five. The reasons for this are complicated, but seven categories provide a closer approximation to the normal distribution.

Figure 4.24: Semantic Differential Question

1. **Indicate the number of times per week you usually engage in technical communications with colleagues in your group.**

1. ☐ (Few)
2. ☐
3. ☐
4. ☐
5. ☐
6. ☐
7. ☐ (20 or more)

Semantic differentials are very useful when the evaluators do not have enough information to anchor the intervals between the poles. However, three major problems detract from this format. First, if the questions are not written with great care, many respondents will not answer or will answer with errors. Second, respondents may flounder and make judgment errors because the semantic differential has no midrange or intermediate anchors. Third, the results lack a certain amount of credibility because they are not tied to a factual observation. For example, compare a factually anchored scale point with a simple enumerated scale point. We find there is a big difference between saying that 70 percent of the respondents said their streams were polluted to the point at which most aquatic life was declining and saying that 70 percent checked 5 on a scale of 1 to 7.

Intensity Paired-Comparison Scales

Intensity scales are very versatile and are sometimes combined with other types of scales. One such combination of scales is sometimes used in establishing priorities. Here an intensity scale is combined with a paired comparison scale. As its name implies, a paired comparison scale compares all the question options by pairs by asking the respondent to rank one item of the pair over the other. An intensity paired comparison scale asks the respondents to scale the amount of the difference between the two pair items. See figure 4.25.

Figure 4.25: Intensity Paired Comparison Scale

Comparison Activities	Much less important (1)	Somewhat less important (2)	Equally important (3)	Somewhat more important (4)	Much more important (5)
1. Biotechnology vs. Acquisition					
2. Description vs. Breeding					
3. Enhancement vs. Preservation					
4. Acquisition vs. Description					
5. Preservation vs. Biotechnology					
6. Breeding vs. Enhancement					
7. Biotechnology vs. Breeding					
8. Description vs. Preservation					
9. Enhancement vs. Acquisition					
10. Acquisition vs. Breeding					
11. Preservation vs. Acquisition					
12. Breeding vs. Preservation					
13. Biotechnology vs. Enhancement					
14. Description vs. Biotechnology					
15. Enhancement vs. Description					

Chapter 5

Avoiding Inappropriate Questions

To make sure questions are appropriate, the evaluators must become familiar with respondent groups—their knowledge of certain areas, the terms they use, and their perceptions and sensitivities. What may be an excessive burden for one group may not be for another. And what may be a fair question for some may not be for others. For example, in a survey of the handicapped, those who were not obviously handicapped were very sensitive about answering questions.

This chapter discusses nine types of inappropriate questions and ways to avoid them. Questions are inappropriate if they

- are not relevant to the evaluation goals;
- are perceived as an effort to obtain biased or one-sided results;
- cannot or will not be answered accurately;
- are not geared to the respondent's depth and range of information, knowledge, and perceptions;
- are not perceived by respondents as logical and necessary;
- require an unreasonable effort to answer;
- are threatening or embarrassing;
- are vague or ambiguous; or
- are unfair.

The best way to avoid inappropriate questions is to learn about the respondent group, design and field test for this group, and not rely on preconceptions or stereotypes. An anecdote may bring this point home. A researcher was pretesting a questionnaire on people who used mental health services. During the test, the researchers expressed surprise that the respondents could handle certain difficult concepts. Annoyed, one of the respondents rejoined, "I may be crazy, but I'm not stupid."

Questions That Are Not Relevant to the Evaluation Goals

A questionnaire should contain no more questions than necessary. Questions that are not related to the goals of the evaluation or that are not likely to be used in the final report should be avoided. They require unnecessary time and effort from respondents. And questions that they view as irrelevant to the evaluation are less likely to be answered. This is the single biggest cause of nonparticipation. However, there are occasions when questions that are indeed very important appear to be irrelevant. If this is expected, the author should be very careful to explain why it was included.

Occasionally, however, someone asks the evaluators to include what is called a "rider"—an unrelated question for use in another evaluation. Including riders creates three problems. First, the evaluation now has a dual purpose that has to be explained to readers. Second, the riders have to be woven into the questionnaire so that they do not seem irrelevant. Third, the use of the rider changes the context and hence the meaning of the questions.

Aside from riders, there are three other ways in which irrelevant questions typically find their way into evaluations:

1. The evaluation design was inadequate. The evaluators did not formulate the overall project questions and the technical approach in a systematic way but decided to measure "everything" and see what they could come up with.

2. The evaluators had a hidden agenda. The evaluation was just a pretext for measuring other things.

3. The evaluators used the questionnaire to cover their bets. They already had the information they needed. They just wanted to be sure not to miss anything.

Not one of these reasons is acceptable because the use of evaluations for such purposes wastes the agency's and the respondents' time and money.

Unbalanced Line of Inquiry

Evaluators should not write questions that could be seen as developing a line of inquiry to support a particular position or preconceived idea, possibly at the expense of evidence to the contrary. The purpose of questionnaires is to develop information for an objective evaluation. To seem to do otherwise threatens a study's reputation for objectivity, commitment to balance, and integrity.

Questions That Cannot or Will Not Be Answered Accurately

Perhaps the most frequent source of error is asking questions that cannot or will not be answered correctly. For example, we asked companies for 4 years of data, when they kept records for only 3 years.

A more difficult problem occurs when respondents either purposely or unconsciously give biased answers. For example, unit commanders had a favorable bias when reporting on the performance of their units, whereas enlisted personnel were more likely to "tell it like it is." Similarly, physicians in certain hospitals rated the quality of their own medical practice very high but were objective in their judgment of peers. In these instances, it was inappropriate to ask unit commanders and physicians to rate themselves, because they were understandably biased in their answers. We obtained much more accurate observations from other sources (enlisted members and physician peer and nurse reports).

Sometimes respondents provide misinformation because they make a random guess or they do not like to admit that they do not know something or they like to please the question asker by responding "yes." But it is better to have no information than false

information. So it is important to skip out those not qualified to answer by using socially acceptable skip questions (see figure 5.1) or to direct the questionnaire only to those the evaluators know are knowledgeable. For example, in one project we evaluated the usefulness of a congressional report that analyzed federal funding by program and geographic location. We did not know which congressional staff used this report. So we analyzed staffing patterns and sent the questionnaire to the right people.

Figure 5.1: Skip Question

1. Was your rating changed by officials other than your supervisor? *(Check one.)*

   1. ☐ Yes (CONTINUE)
   2. ☐ No (CONTINUE)
   3. ☐ Don't know (GO TO QUESTION 21)

Another means of selection is to ask people to rate their expertise. For example, in a study of the feasibility of a national health plan, we asked people to rate their expertise in the various knowledge areas such as the health care industry, insurance, education, manufacturing, and preventive medicine.

Questions That Are Not Geared to Respondent's Depth and Range of Information, Knowledge, and Perceptions

To avoid questions not properly geared to the respondents, it is important not to use words or terms they do not understand. It is very easy to assume that respondents know the same words we do. Some terms and abbreviations that have caused problems in past surveys are "detoxification," "EEO," "DCASR," "peer group," "net sales," and "adjusted gross income." We could have saved time and money had we provided a few words of explanation, such as "detoxification, or drying out"; "peer group, or the people you work with who have similar rank or status"; and "net sales, or the profit on sales after all expenses have been deducted."

Evaluators must also use terms in the same context and sense that people are used to seeing them in. To students at a state college, the student union was a place where people hang out, watch television, and buy coffee and doughnuts; however, to military academy cadets, it was a subversive organization. In another survey, the term "margin" had different meanings to different respondents. It meant barely adequate to consumers, the amount of collateral required for stock purchases to bankers and brokers, the benefits of building or buying additional units to businessmen, and a cross-tabulation calculation to statisticians.

Question writers must be familiar with their population, and they cannot assume too much or too little. For instance, we were worried about using two technical terms in surveying ranchers: "actual grazing capacity" and "forage productive capacity." However, our pretests showed the ranchers uniformly understood the terms. In another survey, we asked users to rate the quality of the computer image tapes from the LANDSAT earth-orbiting satellite. (The tapes provide data used to make computer maps of the earth's surface.) In general, the users could not answer this question because it was too broad. They

wanted us to be much more specific and ask about the quality of the calibration, striping, formatting, wave length bands, pixil resolution, number of original amplitude steps used in digital conservation, corrections for geometric errors and distortions, and threshold settings. In yet another evaluation, we asked state child development and welfare service officials to rate the usefulness of information provided by major federal and state demonstration programs. We found that while the officials could answer for federal programs and for their own state programs, they could not answer for other state programs.

As the preceding examples demonstrate, it is just as easy to assume too much as it is to assume too little. Evaluators usually have to test to be sure. In a survey of welfare recipients, we asked about the difference in quality of service provided by federal government personnel as opposed to state and local personnel. However the respondents saw all as "government men." In another evaluation, we asked mathematics and science teachers to add up a few numbers and calculate some percentages. We assumed this population would have little trouble with simple arithmetic. This was a big mistake.

It is also important to make sure that the question writer's perceptions match those of the respondent's. People from rural areas when asked about a very large company may envision a firm with 50 people and $1 million in sales. Hence, the question writer may want to specify "a very large firm (a firm the size of General Motors, which does several billion dollars in sales and employs more than half a million people)."

Questions That Respondents Perceive as Illogical or Unnecessary

A line of questioning that does not appear to be logical or necessary may tend to confuse or disturb respondents. Questions should proceed in the logical order set up by the instructions and clearly denoted by headings and lead questions. (This is discussed further in chapter 11.) The questions should go from a general topic to the specific item or from the integration of specific details to a logical summary question. Like things should be grouped together, and parts should be structured in a logical progression of function, process, and chronology. For example, a survey of training programs might naturally start with questions on training objectives and then proceed to training plans, curriculums, course programming, lesson plans, instructor selection and training, course material, student selection, student progress assessments, and evaluation. It would, for example, be unnatural to start with evaluations.

Items should not only be logical and relevant but should also appear so. For example, in a survey of postmilitary employment, we were interested only in the major economic sectors likely to do business with the Department of Defense. However, we had to include all major sectors and group these sectors in accordance with Bureau of Labor Statistics classifications, because many respondents were used to seeing the information this way.

Questions That Require Unreasonable Effort to Answer

Evaluators should avoid asking questions that require unreasonable amounts of time or work to answer. In general, it is a good idea to refrain from questions that require extensive and difficult calculations, excessive documentation, difficult to follow and burdensome response formats, extensive analysis and record searches, and a great deal of additional help. "Unreasonable" is a relative term that takes into account what respondents are willing to do, what is fair to ask of them, what the question writer is willing

to do to help them, and what benefits they will get from participation.

In general, form-completion time should be kept to under a half hour. This can be exceeded by a considerable margin if the issue is very salient to respondents; the form is logical, easy to read, and well designed; the approach is right; and respondents both see the need for and value of the information and can reasonably conclude that the evaluators have done all they can to keep the burden down.

For example, we had to divide a very lengthy survey on housing grants into several parts and administer each part to separate individuals so that no respondent had to spend more than 1 hour on the questionnaire. However, in a survey of area agencies on aging, respondents were not the least bit reluctant to devote an entire day to the survey because they felt it was important to their jobs to participate.

Regardless of how long it will take to fill out the form, the writer must be candid about it and tell respondents at the outset how long it is likely to take. Pretesting is the only sure way to find out the completion time, the task burden, the difficulty of the questionnaire, and the respondent's willingness to accept the burden. The price is very high for a miscalculation. Underestimating the burden may increase the nonresponse rate, yield inadequate answers, and lose credibility. If evaluators overestimate the burden, they may unnecessarily compromise the design to gain the acceptance of its users.

Complicated response formats can also be very burdensome. Evaluators should avoid spreadsheet layouts that extend across the page and require respondents to make cross-sectional visual locations. Layouts that make respondents go back and forth

through several pages, learn and remember several difficult codes, and make complicated interpolations should also be avoided.

Threatening or Embarrassing Questions

Questions that are embarrassing, threatening, personal, or sensitive should be avoided. Respondents should not be asked to disclose legal actions or sensitive medical or financial information. Questions should not ask about behavior that makes them look less than ideal or about personal problems. If it is necessary to ask questions of this nature, it should be done in a way that makes the respondents at least minimally comfortable.

For example, in a child-care needs assessment survey, we wanted information on marital status. This question was sensitive because some of the parents had never been married. We collected the information anonymously and explained how it would be handled and used. We expanded the range of the sensitive response category as far as possible without compromising the use of the data. Hence, the marital status choices were (1) married and (2) separated, divorced, widowed, never married. (Approaches for dealing with sensitive questions are presented in more detail in chapter 9.)

Vague or Ambiguous Questions

Vague or ambiguous questions tend to leave respondents frustrated and uncertain how to answer. Vagueness and ambiguity may result from a number of causes, chief (and most remediable) among which are the following four: (1) the writing is unclear, (2) the response choices are unclear or overlapping, (3) the request is not properly qualified, or (4) the question refers to concepts that are too abstract. Unclear writing is covered in chapter 6 and overlapping response choices are covered in chapter

7. This section focuses on qualification and abstraction.

Improper Qualification

Improperly qualified requests do not adequately specify the conditions or the observations evaluators want respondents to report on. If evaluators ask a report user if a report was "timely," the user may not know if they are asking whether getting it took too long or it arrived after it was needed or both. Improperly qualified items are a major source of frustration. Question answerers are frustrated because they do not know how to respond, and question askers are frustrated because they get either no answers or answers they may not be able to use. Some guidelines for correcting this type of flaw are presented below.

First, get to know how the respondent population talks, thinks, and does things. Second, make sure that all terms are well qualified. Third, certain subjects cause problems if they are not part of a person's routine or if their meaning varies with the respondent's perspectives. Some of these subjects are processes, sequences, sources, times, goods and services, organizations, classifications, functions, disciplines, regions, programs, systems, space, business, government, and infrastructures. Fourth, question writers should substitute concrete terms or examples for abstract concepts. Fifth, make as few assumptions as possible.

In a wage and salary survey, we asked business managers to report on their own establishments. We took for granted that everyone would know what their establishments were. However, in these days of chains, branches, decentralized and consolidated offices, holding companies and subsidiaries, this assumption was false. After a few weeks of testing, we finally came up with the following qualification:

"While most of the terms in this questionnaire will be clearly understood, the term 'establishment' may be ambiguous to some and should be further qualified. For this questionnaire, an establishment should be considered as follows:

"A single physical location where one or predominantly one type of business or activity is conducted in your metropolitan area (for example, a factory, store, hotel, airline terminal, sales office, warehouse, or central administrative office).

"Exclude activities that are conducted at other locations, even though they may be part of the business.

"If the establishment engages in more than one distinctly different line of activities or businesses at the same location, consider only the activity that involves the largest number of white collar workers.

"If the personnel office is separate from the business location or serves more than one business, consider only the single separate location in the metropolitan area employing the largest number of white collar workers."

In another survey of personnel, people had trouble answering "Would you relocate?" because they did not know whether we were asking about relocation within the city, within the state, out of state, to the West coast, or to Washington, D.C. Or again shippers could not answer "How many tons of goods did you ship during your last fiscal year?" Goods have different shipping measures: short tons, long tons, tonnage (a measure based on the displacement of water), hundredweights, cords, board feet, cubic feet, cubic yards, and gallons. Finally, while testing a questionnaire in inspector general offices, we were surprised to find that much of the staff lacked audit experience. This was because some of the inspectors general did not consider investigations and inspections as audits. The question should have read, "How many years of experience have you had with

the government doing audits, investigations, or inspections?"

Abstract Concepts

Abstract concepts, like inadequately qualified terms, can be inappropriate because the respondent will have trouble giving a precise answer. Examples are "Does the child-care staff show affection and love toward the children?" "How good was the presentation?" "Do you have sufficient autonomy?" "Assess the neighborhood stability." Respondents cannot readily describe or quantify their observations of love, goodness, autonomy, or stability.

In general, there are four ways to make abstract concepts easier to address:

1. present the concept as behavior,

2. provide definitions that are more concrete,

3. analyze or break out the concept into more elemental and concrete factors, or

4. define the various factors that govern the concept.

The question "Does the child-care staff show affection and love toward the children?" can be broken down into a series of behavior-oriented questions that measure the number and length of times the average child sat on an adult's lap or was picked up, cuddled, or held. Another example of using this behavioral technique is taken from a study of role ambiguity at the U.S. Naval Academy, where the lower-class midshipmen receive much of their training from upper classmen. See question 5.2.

Figure 5.2: Behavior-Oriented Question

1. To what extent, if at all, did you usually know what the upper classmates expected of you? *(Check one.)*

   1. ☐ To little or no extent
   2. ☐ To some extent
   3. ☐ To a moderate extent
   4. ☐ To a great extent
   5. ☐ To a very great extent

Sometimes concepts can be handled more easily by providing concrete definitions. In a survey of program managers, we simplified the abstract question "How much autonomy do you have?" by asking, "How much influence do you have over the project management decisions?"

It may take a lot of work to reduce the abstraction in what appears to be a very simple request. The answers to "How good was the presentation?" may be a composite of many factors. In one evaluation, we had to enumerate these factors and then ask respondents to rate each one. In this case, respondents rated relevance, focus and scope, educational contribution, delivery, planning and organization, and technical merit. Furthermore, the abstractions in these terms had to be reduced by giving concrete definitions. For example, "relevancy" was defined as timeliness, importance, and utility of information, and "focus and scope" were defined as appropriateness of the coverage and the emphasis and detail given to high- and low-priority information.

"Neighborhood stability" was another seemingly simple concept that required substantial explanation. We provided an operational definition of the various factors that governed "neighborhood stability" and asked respondents to rate the extent to which the neighborhood changed with respect to these factors. The factors were new people coming in, residents leaving, new commercial construction, housing construction, housing renovation, number of blighted houses, and proportion of families with children, among others.

Unfair Questions

While irrelevant, unreasonable, embarrassing, threatening, and improperly qualified questions are also unfair, this section focuses on four other kinds of questions that give problems to respondents. These are questions that expose respondents to risk, unnecessarily ask for proprietary information, excessively test a respondent's competence or capability, or entrap the respondents.

We should try to avoid lines of inquiry that put respondents at risk. Examples include asking user groups to report on their regulators, asking employees to report on their management, and asking job candidates to report on merit system abuses. However, sometimes these types of questions must be asked because they are the best or only source of information. When this occurs, the evaluators should be careful to safeguard the respondents' identities and try to prevent any administrative or other uses that would have repercussions on the informants.

For example, we found that certain group homes might be at risk if the information they provided were cross-referenced with that from zoning officials, so we corroborated their reports using other methods.

Evaluators should not ask for proprietary or restricted information unless it is essential to the evaluation. By "proprietary," we mean information on new products, advanced designs, marketing strategies, and so on. Also, restricted information should not be requested, such as data on compliance hearings, equal employment opportunity cases, finances, and national security. Evaluators who need this information should initiate safeguards and maintain a resolve not to disclose it.

Questionnaires that seek to make an audit point by discrediting respondents' capabilities should be avoided. Questionnaires that are the equivalent of an intelligence test or a comprehensive examination of respondent qualifications are unfair. If a competency assessment is necessary for the evaluation, questions can ask about background, achievement, and behavioral information without asking respondents to "take a test."

Evaluators should also avoid using questionnaires for administrative or entrapment purposes—that is, getting respondents to disclose self-incriminating information that may be used against them. Evaluators who must gather this information should be candid and tell the respondents the information they provide might be used against them.

Chapter 6

Writing Clear Questions

To help respondents understand a questionnaire, the question writer must write clearly and at the respondent's language level. The questions must be direct, orderly, precise, logical, concise, and grammatically correct. They must have unity, coherence, and emphasis. Although a detailed discussion of clear writing is beyond the scope of this paper, this chapter discusses some common writing problems and presents general guidelines for increasing the readability of questionnaires.

Simplify the Word Structure

One of the most effective ways to increase readability is to simplify the word structure. Four word structure factors affect readability: the length of a word, the number of syllables in a word, the ratio of root words to words with prefixes and suffixes, and the frequency of a word's use.

Word length should average about 6 letters for the fifth-grade reading level. Sentences with words averaging 10 letters or more are difficult to read.

Cutting back multisyllable words also increases readability. When no more than 8 percent of the words in a sentence have more than three syllables, the sentence is easy to read; when 20 percent of the words have more than three syllables, reading will generally be quite difficult for many respondents. For reading at the sixth-grade level, the average number of syllables per word should be kept under 1.3; for college-level reading, 1.7.

A text is also difficult to read if the ratio of root words to words with prefixes and suffixes is only 2 to 1. Reading becomes easier as this ratio increases. Having four times as many roots as prefixes and suffixes makes for easy reading.

Finally, words that are not in common use are not as likely to be known by people at lower reading levels. Lists and dictionaries that match words to reading levels can be used for assistance.

If the evaluator suspects that readability may be a problem, it should be tested. Several readability indexes focus on word length, number of syllables, word prefixes and suffixes, and sentence length. Examples are the Flesch reading ease formula, the Flesch scale, the Fog index, the Dale-Chall formula, FORECAST, and the RIDE formula.

Be Careful About Words With Several Specific Meanings and Other Problem Words

Sometimes a question is misunderstood because a word in it has several meanings and its context is not clear. For example, evaluators may assume "How significant was that result?" means "How important was that result?" But methodologists may think the question deals with the statistical certainty of the result.

Evaluators who try to improve the readability of questions by using more familiar words often use words with multiple meanings. Some examples are "case," "run," "feel," "fair," "direct," and "line." The question "How many cases do you carry in a month?" will have one meaning to a parole officer and another to a baggage handler.

Other problem words include "like," "best," "believe," "all," "none," "any," and "could." For instance, "like" depends on its context for meaning. Respondents reading "manufacturers like items" may interpret it to mean "manufacturers prefer items" or "manufacturers' similar items." The word "best" can also cause confusion. There is only one best way, but how often do questionnaires state, "Check all the answers that best apply"? The word "believe" may mean "think" to some and "have a conviction" to

others. Because "all" and "none" are absolute words, people who are quibblers may avoid these words, insisting there are no such absolutes. "Any" can mean every or some. And "could" is often confused with "would" or "should."

Do Not Use Abstract Words

Abstract words, or words that convey general or broad meanings or relationships, should be changed to concrete words, or words with more specific meanings. Concrete words are more easily understood. Consider the following example: "Enumerate the mishaps attributable to personnel not cognizant of the regulation that could have been obviated." After we replace the abstract words with concrete words and reorganize the sentence, it becomes much more easily understandable as follows: "List the preventable errors caused by people unaware of the regulation."

However, an undue emphasis on concrete words may cause an overly detailed, inefficient line of questioning. It is important to choose the appropriate level of abstraction. As a rule of thumb, the lower the expected level of a respondent's literacy, the more concrete the words must be.

Reduce the Complexity of Ideas and Present Them One at a Time in Logical Order

Question writers must be concise because they need to cover a lot of topics with as few questions and words as possible. However, they sometimes defeat their own purposes by too quickly presenting complex ideas and by failing to link the ideas in logical order. For instance, consider the following question, addressed to National Guardsmen and Reservists:

"What percentage of your mission training and the occupational specialty training that you received during unit assemblies and annual active duty followed a published training schedule?"

A less complex, more logical version of this question might read:

"We need to know what proportion of your training followed a published schedule.

"First consider the mission training you received during the unit assembly. What percentage of this training followed a published schedule?

"Next consider the mission training received during annual active duty. What percentage of this training followed a published schedule?

"Now forget mission training and concentrate on military occupational specialty training. Consider this training received during the unit assembly. What percentage of the occupational specialty training followed a published schedule?

"Finally, consider the military occupational specialty training received during annual active duty. What percentage of this training followed a published schedule?"

Reduce the Sentence Length

Sentence length has a big effect on readability. Longer sentences usually contain more information, are grammatically more complex, and are harder for the reader to process. It is a rule of thumb that 10-word and 11-word sentences are suited to a sixth-grade reading level. Every two or three words added to a sentence, up to a 16-word sentence, increase the reading level by about one grade. After this, every word increases the reading level by one year. Hence, sentences of 25 words or longer may require college reading levels.

Simplify the Sentence Structure

One factor that makes question writing difficult is the need for very precise, well-qualified language. To satisfy this requirement, sentences grow in length and become more complex. Although the effects of syntax

on readability are not well understood, complex syntax also appears to be associated with reading difficulty. However, as we explain in the next paragraphs, this may result more from a tendency to bury, or embed, a main idea in complex sentence structure than from a problem with complex sentence forms.

The simple sentence, containing a clear subject-verb relationship, should be the writer's goal. However, because of the need for modifiers, qualifiers, and variety, more complicated sentence forms will have to be used at times.

Here are some rules of thumb. In a complex sentence, the main idea should be at the beginning. If this is not possible, it should be at the end. Embedding the main idea in the middle of the sentence should be avoided. The number of dependent clauses should be limited. Compound sentences should not be used unless the independent clauses are of equal value. Otherwise, the less important clause will take on undue importance. As for compound-complex sentences, they should be avoided, if possible.

Use Active and Passive Voice Appropriately

People read faster with more comprehension when the text is in the active voice than they do when it is in the passive. In active voice sentences, the emphasis between the subject and verb is clear and the action moves smoothly. Nevertheless, in question writing, certain thoughts should be emphasized more than others. The passive voice can be very useful in subordinating the subject or focusing attention on the object in the sentence.

Use Direct, Periodic, and Balanced Styles Appropriately

Most questions should be asked in a direct style with the main thought first and the details and qualifiers later. This form, sometimes called a "loose sentence," allows quick development of the main idea and the addition of details without the confusion caused by embedding. However, the question writer should be careful not to dilute the main idea by overloading the sentence.

Sometimes the "periodic style," in which the main idea comes last, is more useful. For example, when a complex idea must be expressed in one sentence, the writer can build up or emphasize the thoughts the respondent must consider.

On occasion, evaluators may present the reader with a balanced contrast of two equal ideas. When this occurs, the two ideas are presented in like construction.

Avoid Writing Styles That Inhibit Comprehension

Question writers should avoid needless shifts in subject, person, voice, and tense. Wordy writing styles should also be avoided. Cutting down on the number of words and sentences allows the respondent to focus more on the information being presented. Concise writing can also add force and emphasis to a query.

Prepositional decay is a serious problem in question writing. It often develops in the simple sentence, in which the writer adds so many qualifiers that the main idea is diluted, deemphasized, or forgotten. Although not as serious a problem as embedded syntax, it can compromise a question's effectiveness. Here is an illustration of prepositional decay and a simplifying revision. Prepositional decay: "The federal government, which has a number of programs to provide assistance to individuals and public and private organizations through the state and local

governments for use in planning, implementing, and evaluating housing activities in community development areas, is consolidating these categorical grants under a single block grant." Simplifying revision: "The federal government is consolidating its categorical grant housing programs into a single grant. This grant, called a 'block grant,' can be given to a state or local government."

Repetition and parallelism can aid comprehension. However, when overused, these techniques become monotonous and irritating.

Because people generally have more trouble with an idea stated negatively, question writers should avoid negatives. It takes longer to read negatives and they make for more mistakes. These problems are exacerbated when double negatives are used, even though they may be logically correct.

Although researchers are not quite sure why, they have found that another readability problem develops when writers create a noun from a word that is normally a verb. For instance, the nouns "specification, "participation," and "implementation" were derived from the verbs "specify," "participate," and "implement." Rather than adding a level of abstraction that slows the reader down, question writers should go back to the original verb.

Often, seemingly small mistakes can cause a lot of trouble. Misplaced modifiers, for example, confuse the reader. Pronouns are sometimes placed where their antecedents could be more than one word. On occasion, the reverse occurs, and the antecedent of the pronoun is made vague or indefinite or put in the wrong position. A similar problem arises when the word "which" is used to refer to a clause. The clause is perceived as indefinite and the reader is confused.

If the clause cannot be reduced to one word, the sentence should be reworked to eliminate "which."

The following question has a similar problem: "If you do not have children younger than 12 living with you now, is this likely within the next 2 years?" Because the antecedent of "this" was unclear, some people thought that a "yes" answer meant that they did not have children younger than 12 living with them now and did not expect children to be living with them in the future. However, others thought that a "yes" meant that they expected to have children living with them within the next 2 years. A better way to ask for this information is to ask two questions: "Do you have children younger than 12 living with you now?" and "Do you expect to have children younger than 12 living with you within the next 2 years?"

Chapter 7

Developing Unscaled Response Lists

A type of multiple-choice question known as an unscaled response list is frequently used in GAO questionnaires. We develop a list of entries and ask respondents to select one or all that apply. In some instances, we want respondents to rate each category for degree of importance or satisfaction.

To prepare a good unscaled response list, the question writer must have a thorough grasp of the subject matter covered by the question and understand the subject from the respondent's perspective. Only then can unscaled response lists meet the following standards:

- The lists must contain all the categories perceived by respondents as significant to the question topics.
- The categories must not overlap.
- The categories must be relevant and appropriate from the respondent's perspective.
- The lists should not exceed five to nine categories, unless the categories are grouped into sets.
- The specificity of the response categories must be at the level of detail required to answer the evaluation question.
- Respondents must feel that the order in which the categories are presented is logical.
- A prior screening question should be used if the question does not apply to all respondents.

Developing Comprehensive Lists

To obtain useful data, response lists must contain all important categories that apply to the question area. Usually, the question writer includes an "other (specify)" category to cover omitted alternatives. However, because respondents are more likely to recognize than recall all the factors they want to report, they tend to underuse the "other" category, therefore omitting an important alternative.

Do your research; write your list; then pretest. In most cases, pretesting is invaluable for ensuring the adequacy of the response list because the respondent population usually knows the area better than the evaluators do. Seemingly broad topics like the quality of medical care can be resolved into comprehensive lists through research and analysis. For example, consider the question in figure 7.1 used successfully in evaluating veterans' satisfaction with Agent Orange examinations provided by Veterans Administration medical centers.

Figure 7.1: Question With Comprehensive List of Categories

1. Which, if any, of the following laboratory tests were given to you by the VA as part of your Agent Orange examination? *(Check all that apply.)*

   1. ☐ Blood sample
   2. ☐ Urine specimen
   3. ☐ Chest x-ray
   4. ☐ Other x-ray
   5. ☐ Sperm sample
   6. ☐ Skin sample
   7. ☐ Other (please specify)

Presenting Mutually Exclusive Categories

To develop nonoverlapping categories, the question writer should use words that clearly define category membership. For example, to determine the marital status of respondents, the writer should avoid using the separate categories "single" and "divorced or separated." The word "single" can be read as applying

to either divorced or separated as well as never married people. Another example of overlapping categories is given in figure 7.2.

Figure 7.2: Question With Overlapping Categories

2. What is your occupation? *(Check one.)*

1. ☐ Manager
2. ☐ Professional
3. ☐ Technician
4. ☐ Secretary
5. ☐ Sales person
6. ☐ Other (specify) _____

Because the categories in figure 7.2 are not sufficiently qualified, they are not mutually exclusive. In particular, managers, technicians, secretaries, and sales persons all consider themselves professionals.

Several techniques can be used to develop number ranges that are mutually exclusive. For example, adding such text as "less than 6 months" and "from 6 months up to a year" helps respondents answer questions involving time. In a question about a respondent's age, the end points of one response category must not overlap the beginning of the next category. See figure 7.3.

Figure 7.3: Question With Nonoverlapping Categories

1. What was your age when you filed your bankruptcy petition? (For joint cases, check age of major wage earner.) *(Check one.)*

   1. ☐ Under 25 years of age
   2. ☐ 25-34 years of age
   3. ☐ 35-44 years of age
   4. ☐ 45-54 years of age
   5. ☐ 55-64 years of age
   6. ☐ 65 years or older

Sometimes a question focuses on two or more information items rather than one, causing overlapping categories. For example, we wanted to know how former Department of Defense employees had learned about postemployment restrictions. The word "how" in this context has various meanings: from a co-worker, at a retirees' meeting, at the office, from magazines or newsletters, during an exit interview at the department, and so on. A response list with these options would be confusing, because it mixes sources of information and places of learning the information. Rather than asking "How?" we needed to ask either "From whom did you learn . . . ?" and "Where were you when you learned . . . ?" or, better still, "From what source did you learn . . . ?"

Using Relevant and Appropriate Categories

The alternatives provided in a response list must be geared to the respondent group. For example, if we are surveying food stamp recipients, the response categories for a question on yearly income should be

skewed toward the low end of the income range. If we provide response alternatives of $0 to $10,000, $10,001 to $20,000, and so on, most if not all the respondents would probably select the $0 to $10,000 alternative, and the data would not be very useful. A more appropriate format would be $0 to $2,000, $2,001 to $4,000, and so on.

To write relevant and appropriate items, the question writer should tailor the wording to the majority of respondents. An illustration is in the use of medical terms. If we need to measure the receipt of health services, we might use simple terms and give examples, if the respondent is not a medically trained professional. See figure 7.4.

Figure 7.4: Tailored Question With Comprehensive Nonoverlapping Categories

1. What services, if any, have you received from your health maintenance organization in the past calendar year? *(Check all that apply.)*

   1. ☐ Surgical services
   2. ☐ Medical services for conditions of the bones, muscles, and tendons, such as breaks, strains, or sprains. In other words, orthopedic services
   3. ☐ Eye care, diagnosis, or treatment: Ophthamology
   4. ☐ Ear, nose and throat care: ENT
   5. ☐ Mental health or psychiatric service
   6. ☐ Arthritis or rheumatism treatment
   7. ☐ Allergy
   8. ☐ Other

Keeping the Response List Reasonably Short

People can focus on lists of about five to nine categories. Longer lists should be grouped into sets with titles to help respondents grasp the range of information. When each of the response categories is to be rated (for example, by degree of importance), subgrouping also aids respondents in assessing each entry's relative value. Long response lists are more subject to primacy and recency effects. If respondents are asked to select entries from a long list, they tend to select the first and last entries. (Primacy and recency effects are discussed in chapter 8.)

Using Categories of Appropriate Specificity

Response categories should be neither too broad or abstract nor too narrow or specific for the measurement purposes, and the specificity should be tailored to each respondent group. The level of response specificity also sets that of the question stem, which should be at one level more abstract than the response space. To measure the quality of a speech, for example, we might ask people to assess its educational value, focus and scope, clarity of delivery, interest value, and topic emphasis. Each of these categories is appropriate to the level of measurement needed for evaluation. More specific information on the clarity of the delivery through diction, accent, and syntax is more detailed than needed to answer the evaluation question.

While it is important not to ask for details you do not need, it is just as important to ask for levels of detail that you do need. A survey on water pollution further illustrates this point. When the Environmental Protection Agency asked paper-manufacturing plants about the acidity and alkalinity (pH) of waste water released into rivers, the response categories were not precise enough. The agency asked whether the pH level was 4 to 5, 5 to 6, 6 to 7, 7 to 8, and up but needed to know whether the pH level was 7 (6.5 to 7.4), which is neutral. A pH scale of 6 to 7 includes measures that are acidic. A pH scale of 7 to 8 includes measures that are alkaline.

It is also important that the level of specificity meet the expectations for the target population. For example, in a national parks survey of people who burn wood, the question and responses varied in specificity according to the knowledge of the types of people cutting the firewood. For the usually less knowledgeable fireplace users, they ask "what types of woods do you usually burn?" with answer responses of hardwood, softwood, mixed wood, any type of wood, and don't know. But for the more

knowledgeable wood stove users, they ask "what types of hard and soft woods do you usually burn?" with answer responses "locust, oak, cherry, hickory, pine, poplar, and cottonwood."

Listing Categories in the Logical Order Expected by Respondents

When respondents read a question, they begin to anticipate the response alternatives. If the alternatives are presented in a sequence that is not perceived as logical, the respondents may feel they have misunderstood the question and return to study it again. (Logical sequence is discussed in chapter 11.)

Using a Screening Question

Response lists may place an implicit demand on respondents to check an entry. For example, if doctors are asked to report the professional publications they read during a 2-week period and are presented with a list, they will probably check something regardless of whether they have read a journal or not. Using a screening question that asks whether or not they had been able to read any publications in the last 2 weeks would reduce this tendency.

Chapter 8

Minimizing Question Bias and Memory Error

Question Bias

Sources of bias can occur in either the stem of the question or the structure of the response. Various types of biased questions, as well as some ways to avoid them, are discussed below.

Status Quo Bias

Questions that state or imply prevailing conditions may produce inaccurate data. In the following examples, the use of "most" and "as it now stands" could influence answers:

"Most child support enforcement offices confirm the employment of absent parents on a regular basis (such as monthly or every other week) rather than 'as needed' (such as when support payments are not made or when files are transferred). Does your office confirm the employment of absent parents regularly or on an 'as needed' basis?"

"As it now stands, Department of Defense policy is to provide civilian employees with information on postemployment restrictions during exit interviews. Did you receive any information on employment restrictions when you left the department, or did you leave without getting this information?"

Better presentations of these questions would delete status quo information, since some respondents would otherwise feel compelled to conform to what is seen as "normal."

Bias in More Than One Direction

Sometimes question writers add qualifying or identifying information that can bias respondents in different directions. For example, a question writer might ask, "Who would you vote for, Pat Green, the Republican incumbent, or Chris Lamb, the Democratic challenger?" If the question writer is interested in the choice between Pat Green and Chris Lamb, the question is biased. The respondent's choice will be influenced not only by the persons individually but also by political party and the difference between

continuance and change in leadership. An illustration of this type of bias in a GAO study might be the following:

"Should program managers with responsibilities for major weapon systems be civilians with an engineering background or military personnel with an operational background?"

If we want people to base a choice on whether the managers are military or civilian, we must take out the engineering and operational qualifications. If we want people to base a choice on operational and engineering qualifications, we must take out the military and civilian comparison. If, however, we want them to base a choice on several factors, all the factors must be presented. Consider the following example: "How important, if at all, is it for the project managers to be civilians or military or have engineering or operations backgrounds?"

Bias From Specific Words

Certain words are "loaded" because they evoke strong emotional feelings. In our culture, such terms as "American," "freedom," and "equality" may tend to evoke positive feelings and "communist," "socialist," and "bureaucracy" may tend to evoke negative feelings. Other emotionally laden words, such as "abortion," "gun control," and "welfare," probably evoke a complex pattern of responses. Since it is difficult to control or predict the effect of these words, it is usually best to avoid them. We can illustrate phrasing that could bias responses. See figure 8.1.

Figure 8.1: Biased Question

1. There has been a great deal of discussion lately about having the federal government take over the costs of welfare. Which of the following statements comes closest to your opinion? *(Check one.)*

   1. ☐ It is up to the federal government to take care of people who don't work.
   2. ☐ People who don't work already receive enough welfare—the federal government shouldn't provide any.

Phrases such as "people who don't work" do not contribute to an objective frame of reference. (See Warwick and Lininger, 1975.)

An example from a GAO study involves a mail survey of private industry's views on competitive bidding practices for major weapon systems. An article by an expert had compared the bidding process to a game of "liar's dice," implying that bidding is like a game that favors a skilled deceiver. The use of the term "liar's dice" could elicit a negative or threatened feeling. Instead, we wrote the question as follows:

"One approach to bidding might be to be conservative. That is, to overestimate cost and underestimate performance on the theory that a firm will look better when it delivers because it beat its original estimates. Another approach would be to make a realistic bid by specifying the actual costs and expected performance. Still a third approach would be to be optimistic by understating costs and exaggerating performance. You might do this on the theory that if you are not optimistic, you won't get the job. The question is,

Which strategy gives the best probability of winning: making conservative and realistic estimates or optimistic estimates?"

Interestingly, a single word can affect how people respond to a question. For example, people viewing a film that shows a car crash will probably report broken glass if we ask them what happened when the car was "smashed"—even if the film does not show any glass breaking—but they would not report broken glass if we ask them what happened when the car was "hit."

Unbalanced Question Bias

Just as we can have unbalanced response categories (see chapter 4), we can have unbalanced questions. The wording of an item stem or question may imply or suggest how the respondent should answer. "Do you support the establishment of group homes for the mentally retarded in single-family zones?" or "You're the best trained soldiers in the world, aren't you?" might elicit positive answers, since no other possibilities are made explicit. Questions can frequently be balanced by adding "or not" ("Did you get training or not?") or word opposites ("Do you support or oppose?")

It is important to balance word opposites well. For example, "forbid" and "not allow" have different meanings and cannot be used interchangeably as opposite terms for "allow." Depending on the context, "dissatisfied" is the appropriate opposite term for "satisfied," while "not satisfied" is inappropriate. For example, some studies of employee satisfaction indicate that those who are "not satisfied" with their work are basically content but would like improvements in some areas. In contrast, employees who are dissatisfied are basically unhappy with their work.

Questions That Omit Important Factors

The answers respondents give to a question vary according to their frame of reference. For example, some employees might judge their job satisfaction on their commuting time, some on promotion policies, and others on types of tasks and responsibilities. The question asker must ensure a common frame of reference by delineating each of the factors respondents should consider in reaching an answer. This is particularly important when the respondent has a vested interest in the subject and when complex questions containing several aspects are being asked.

Even though a question may be formally balanced, one position may be favored over another because of the topic and the respondent's characteristics. For example, we asked farmers "Do you think the government should provide free agricultural weather reports or not?" Expecting a yes bias, we needed to get the respondents to consider the question from a variety of viewpoints. For example,

"In reality there are no free services or subsidies since ultimately everyone pays taxes to provide them. The question is, Do you favor free weather reporting services even though all taxpayers must bear the cost?"

In a survey question mentioned previously, program managers of major weapon systems were asked whether civilian or military personnel should be program managers. Most of the respondents were military. To obtain opinions based on balanced considerations, we presented the pros and cons:

"A persistent issue is whether or not the program manager [PM] position should be held exclusively by military personnel. There are advantages and disadvantages attributed to the military PM. Pro-military arguments claim knowledge and appreciation of the system (conditions, personnel, organization, etc.) and advantages of service affiliation. However, the military PM system is sometimes criticized for short tenure, valuing performance over

cost, constraints on independent action from the military rank hierarchy, and service-mission suboptimization. The question is, Should the federal PM work force be composed exclusively of military personnel or should it be composed of both qualified military and civilian personnel?"

Broad questions contain many different aspects to be evaluated. People tend to be selective in remembering and consider only some arguments. The question writer should present all the significant factors and should balance the pro and con positions. If three arguments are given in support of a position and two arguments are given in opposition, endorsement percentages will tend to favor the former.

Primacy and Recency Effects

Structured response formats vary in length from two alternatives (such as "yes" and "no") to fairly lengthy lists. The evidence in survey research is mixed regarding the tendency of respondents to pick alternatives presented first (primacy effect) or last (recency effect), regardless of item content.

Primacy effects may result because the first item in a series may receive additional attention or mental processing. Recency effects seem to be more likely when the reader is presented with lengthy or complex text, lists, or response alternatives. When presented with a questionnaire item, people try to process both the question, or the stem part of the item, and all the choices in the response part of the item before answering. Hence, respondents with long lists of response alternatives tend to be biased toward the last few items, because that is the material they have been exposed to just before they are ready to answer.

However, primacy and recency effects often work in tandem. This results in higher reporting for the first and last few choices in a list and lower reporting for the middle items. The effect of these biases is also

dependent on the media used. In self-administered instruments, primacy effects dominate. But the opposite is true for personal interview and telephone interview surveys. For these, recency effects dominate. Some of the best ways to minimize the differential effect is to keep the list short or add subtitles and use formats that present the list in shorter groups. (See figure 8.2.) Another way is to slow the reader down by turning the "check all that apply" format into a "check yes or no" format. (See figure 8.3.)

Figure 8.2: List Divided Into Subgroups to Counter Primacy and Recency Biases

EMPLOYMENT

1. ☐ Unable to work because of illness or accident

2. ☐ Periods of unemployment because of job layoffs, job changes, strikes, seasonal factors, etc.

3. ☐ Cutback in hours worked per week (e.g., loss of overtime; work slowdown; or self-employed, lack of work)

FINANCES AND CREDIT

4. ☐ Loss of second income (e.g., spouse became unemployed)

5. ☐ Unusual medical bills (e.g., doctors, hospitals)

6. ☐ Divorce, separation costs; alimony or child support

Figure 8.3: "Check All That Apply" Response Format Changed to "Check Yes or No" Format

(Check all that apply)

1. ☐ Unable to work
2. ☐ Laid off work
3. ☐ No steady employment
4. ☐ Loss of second family income
5. ☐ Unusual medical expenses
6. ☐ Divorce

(Check yes or no)	Yes	No
1. Unable to work	☐	☐
2. Laid off work	☐	☐
3. No steady employment	☐	☐
4. Loss of second family income	☐	☐
5. Unusual medical expenses	☐	☐
6. Divorce	☐	☐

Bias effects from prior processing of an item or having prior concern with the topic can sometimes be

ameliorated by the placement of the item in the questionnaire. For example, community opposition is frequently cited as a problem in locating group homes in residential areas. In surveying people who operate group homes for the mentally retarded and emotionally ill, we asked them to respond on a five-choice scale. We expected a tendency on the part of respondents to focus on positions they had encountered. Therefore, the scale was constructed with support and opposition for opposite poles. Furthermore, we counteracted the inherent bias by presenting the support anchor as the first unit and the opposition anchor as the last unit. This example is presented in figure 8.4.

Figure 8.4: Using Presentation Order to Counteract Expected Bias

1. Consider the individuals and groups in your community who were contacted. Overall, how did their support and opposition compare? *(Check one.)*

   1. ☐ Expressed **much more** support than opposition
   2. ☐ Expressed **more** support than opposition
   3. ☐ Expressed **as much** support as opposition
   4. ☐ Expressed **less** support than opposition
   5. ☐ Expressed **much less** support than opposition

Presenting Choices in a Logical Sequence

A list of unscaled response alternatives (reasons for going bankrupt, characteristics of grazing land, and the like) must be put in a logical order. That is, the

options that are of primary significance to the topic being considered should be listed first. Otherwise, we will violate a rule of conversational English and perhaps confuse the respondent. For example, a questionnaire asking people why they dropped their memberships in health maintenance organizations would present the ability to choose doctors and the quality of care at the beginning of the list and paperwork and hospital decor at the end.

Use of the "Other" Category and Incomplete Lists

Question writers often include an "other" category in unscaled response lists as a check for the completeness of the lists. The "other" category offers the respondent the opportunity to add the additional salient responses that the writer missed in providing a comprehensive range of choices.

Omitting viable options as well as the other category causes overreporting in the categories presented because the respondents will force the omitted choices into these categories. Similarly, they sometimes overreport in the "other" category for the same reason.

It is essential that the evaluators analyze responses in the "other" category to (1) determine the adequacy of the choices listed and (2) make adjustments for underreporting in the major categories (for example, one respondent wrote "availability of housing" under "other" when availability of housing stock was listed as an entry).

Biased Examples

Sometimes questionnaire writers provide examples to illustrate the kind and range of information needed. Single illustrations may cause a respondent to restrict a frame of reference. For example, were we to ask students how satisfied they are, if at all, with their teachers and mention the name of only one teacher,

we might get their evaluation of only that teacher rather than of their teachers in general.

Memory Error

Many factors affect memory: the time since an event occurred; its saliency; the respondent's motivation, ability, and experience; the type of material to be recalled or recognized; and, most importantly, the way in which the questions and reporting formats are crafted. Memory error can result in either underreporting or overreporting. Memory error is revealed in three ways: omissions (forgetting that an event occurred), intrusions (recalling an event that never occurred), and event displacement (miscalculating when an event occurred).

Consider the dynamics of memory in answering questions. The respondent must comprehend and interpret the question; decide what information is needed; search his or her memory; select, analyze, and integrate the information; and make a judgment on what and how to report. To do all this, the respondent behaves in part as if his or her brain functioned with two types of memories: a short-term memory and a long-term memory. Respondents use the short-term memory to remember the question text long enough to understand and interpret the question and initiate a retrieval process from the long-term memory. People usually retain the short-term memory information only long enough to use it (18.7 seconds). For example, they usually forget the telephone number they have just dialed or the syntax of the question they have just answered.

While the two memory functions work together complementarily, they appear to be quite different. The short-term memory processes information much more quickly than the long-term memory. It handles information in limited sets of about seven chunks or units and stores it as a representational image. The

slower long-term memory retrieval system stores and accesses most of what the respondent knows. It stores this information as semantic or meaningful codes rather than as representational images. We need to understand this difference because to facilitate recall, we need to write questions that satisfy the requirements of both of these processes.

Facilitating the Memory Process

Some practices that have been shown to facilitate the memory process follow.

1. Use simple, direct, organized, and specific language. The memory process is facilitated by using an organized line of inquiry, by using the active voice, by avoiding lengthy qualifications, by using familiar words with a limited range of meanings, and by using simple syntax. (Complex syntactical constructions often embed the main point.) This language style facilitates the short-term memory process, because it allows the respondent to quickly identify the type of information needed for the answer without taxing the short-term processing capabilities. It also helps the long-term process by aiding the respondent to remember.

2. Be consistent with the way people have learned the information they are asked to remember. Present the question material in the same sequence, manner, terminology, level of conceptualization, detail, and abstraction in which the information was learned or is usually experienced. Sometimes even slight changes can interfere with retrieval. For example, it takes the average person 10 seconds to report the months of the year in calendar order but 2 or 3 minutes to report the months in alphabetical order.

3. Avoid reversals. For a variety of reasons, English uses negative subordinations, prepositions, and other language codes to reverse the meaning, order, or

importance. For example, "not unlikely" means "likely," and "performance was worse under PFP than TQM" means that performance was better under TQM than PFP. Avoid these reverse constructions. The memory system sometimes forgets the reversal code, so that the information is recalled incorrectly.

4. Make sure the questions and the reasons for asking the questions are meaningful to the respondent. People are more likely to be able to recall information they believe is important.

5. Use the question to guide the answer search. People remember things better when the topic of the search is specified initially. They store information in related and hierarchical categories. Therefore, the question, or the stem part of the item, should ask the question in a simple, complete, and direct manner, and it should specify the category or type of material to be searched for at a level only somewhat more general than the details or specific choices presented in the response space. For example, consider the following question: "How satisfied or dissatisfied are you with the following components of your benefit package?" This stem is not as effective as an alternative stem that asks "How satisfied or not are you with the following benefit and pension components of your compensation package?" because people tend to see the compensation package as having quite different classes of components—benefits and a pension. Avoid stems that attempt to shorten the question and not identify the search category. For example, "How satisfied or not are you with the following?" Worse yet are one- or two-word stems that imply a question and use the complete sentence format. An example is "satisfaction with: health insurance? life insurance? etc."

6. Do not overload the short-term mental processing system with too many alternatives, considerations, or qualifications. An example is "How satisfied or dissatisfied were you with the information you obtained, if you obtained information, on 'Brassica cultivars' in current use, special genetic stocks, obsolete cultivars, traditional varieties or landraces, distant relatives of cultivated varieties that form fertile hybrids, varietals that can be crossed and varietals that can be crossed with advanced techniques on the amounts of resources existing in nature, and gene banks and the amounts of resources that are in decline?" Here the short-term mental processing system is obviously overloaded because the respondent is asked to consider too much information simultaneously; in such cases, the respondent often resorts to inefficient coding and long-term system access strategies. In the stress of inefficient strategies, the information retrieval, integration, and judgment functions of memory recall are usually the first to break down.

This does not mean that evaluators cannot make complex inquiries. But they must limit the information to be kept in the respondent's head during the comprehension and retrieval tasks to a small number of units that can be immediately processed in discrete steps. The set size of information units that can be kept in one's head varies from a few units to about seven, depending on the similarity and complexity of the units. Complexity can be handled by increasing the number of steps and presenting them serially. For example, consider the preceding example when it is decomposed into the sequence of questions in figure 8.5.

Figure 8.5: Complex Question Broken Into Sequence of Questions

1. Have you obtained information about the amount of genetic resources existing in nature or in gene banks or about the amount of genetic resources that may be declining? *(Check one.)*

   1. ☐ Yes (continue)
   2. ☐ No (go to question 33)

2. If yes, which, if any, of the following types of Brassica genetic resources did you obtain this information for? *(Check all that apply.)*

   1. ☐ Cultivars in current use
   2. ☐ Special genetic stocks
   3. ☐ Obsolete cultivars
   4. ☐ Traditional varieties/landraces
   5. ☐ Distant relatives of cultivated varieties that form fertile hybrids
   6. ☐ Distant relatives of cultivated varieties that can be crossed using conventional methods but with a high level of sterility
   7. ☐ Distant relatives of cultivated varieties that can be crossed using advanced techniques

3. In general, how satisfied or dissatisfied were you with the overall quality of the information obtained? Answer for the amount of resources existing in nature, the amount existing in gene banks, and the amount considered to be in decline. *(Check one column for each row.)*

	Very dissatisfied (1)	Generally dissatisfied (2)	As satisfied as not or undecided (3)	Generally satisfied (4)	Very satisfied (5)	No basis to judge (6)
1. Amount of resources existing in nature						
2. Amount of resources existing in genebanks						
3. Amount of resources in decline						

While the proposed alternative may have more words and structure, overall this line of questioning will

provide faster, less burdensome, and more accurate answers.

7. Make the judgment that the respondent has to make during the retrieval from long-term memory as easy as possible. Recall is much less accurate for information that is complex, multivaried, vague, or with conflicting elements. For example, consider the following question.

"Please provide an overall assessment of all the GAO reports that you read last year with respect to the following considerations: timeliness, clarity, quality of reporting, responsiveness, comprehensiveness."

It would have been better to ask about a specific report that the respondent had read. Then specific questions should have been asked about each of the attributes in which the evaluators were interested. Each attribute should have been carefully specified in concrete, operational, and meaningful terms. For example, timeliness should have been resolved into two components—turnaround time and the provision of information in time to use it. The quality of the reporting should have been given at least four properties: focus and scope of the reporting, the soundness of the evidence provided, appropriateness of the qualifications, and logic of the conclusions. In a case like this, overall assessment questions should be considered last after the respondents have refreshed their memory on all the properties to be evaluated.

8. Use cues to help respondents retrieve data from their memories. Some of the examples below show a wide variety of cuing methods. For example, rather than ask "what are the ages of your children?" ask "Starting with the oldest give me the ages of your children as of their last birthday." The second alternative uses explicit language and time and episode incident referents as well as a natural order

as cues. The following example uses place, time, episode, routine memory, qualification, quantification, and people referent cues: "Think about where and whom you were with when you ate breakfast this morning. List the foods you ate. Did you have more than one serving of any of these foods? Did you eat less than half of any of the servings?" Questionnaire writers also use examples to trigger recall. Consider the following case in point: "So far this year the Suburban Trust Company reported that 10 percent of their windowed return envelopes were incorrectly posted because the mailer inadvertently put the mail insert over the address that was supposed to show though the window. The question is, when was the last time you did that?"

In addition to these more familiar techniques, researchers have developed some rather ingenious way of cuing. Specifically, crime report surveyers found that the conditions and activities that the victims might have experienced reduced memory errors by 30 percent, as in "Think about the times you came home late at night last year. Were you ever robbed or assaulted?"

In still another approach, authors cued recall by varying the respondents' orientation. Respondents were asked to recall the details of a house visit from the viewpoint of a prospective home buyer, then from the perspective of a burglar. Each successive recall produced new and accurate observations.

There are other types of cues. Examples are a calendar of political, newsworthy, or administrative events; a list of names, topics, or events relevant to the material to be recalled; and a narrative description of the respondent's routine.

Question writers also use the questionnaire text to cue memory. Respondents receive cues from the

organization of the survey, the line of questioning, the direction given in the question stem, the instructions, the presentation of the response alternatives, and the emphasis given.

There are, of course, problems with cuing. First, if inappropriate cues or miscues are presented, the respondent usually produces an inappropriate or erroneous answer. Second, respondents who have no memory of the event or truly cannot recall may feel pressured to answer with false reports based on the cues offered rather than saying, "I can't recall." Third, if the cues are leading, the respondents may follow the cue and bias their answers. Fourth, cues that use special terms or difficult words often confuse the respondent.

9. Consider using longer questions sometimes. Longer questions may set the scene by presenting significant aspects of an argument, defining how terms are to be measured in the question, or giving examples. Short questions sometimes achieve their brevity by means of complex words. To say the same thing more simply takes some effort but may reap rewards by increasing a respondent's memory and comprehension.

10. Always consider the respondent's ability, motivation, or viewpoint. Respondents who are fresh, interested, alert, confident, and smart answer with less memory error than those who are fatigued, bored, unobservant, anxious, under stress, or less capable. Respondents who have a strong concern or bias toward a particular issue will remember things that support that viewpoint and forget facts that do not. Material that is well learned, or that the respondents are familiar with or have thought about or been extensively associated with, will be recalled much better than material that has had more limited exposure.

11. Consider the limitation on memory. Unless the information is memorable or well learned or unless a conscious effort is made, people do not store details in their memory. They will forget about 75 percent of the detailed information they were exposed to in 1 to 3 days. They code and store the information as a summary, organized around the essential facts or salient features of the observation, experience, event, or material. After longer periods of time, months and years, they may forget even these summaries. Their memories will sometimes distort or selectively add or subtract or otherwise alter the information stored. If the stored information is inconsistent or not meaningful or rational, their memories will omit the inconsistencies and add material that was never originally stored in their minds to make the information consistent, meaningful, or rational. If they later find out that some aspects of the mentally stored information are inconsistent or not important, they will forget that information and again add new information and correct important old information that was never part of the original memory.

In short, people are likely to remember the gist of an event better than its details. If we need highly detailed information, we should consider using other data-collection sources, such as observations, diaries, and records rather than self-administered questionnaires. If mail questionnaires do require detailed information, respondents should be asked to refer to their records; however, the burden of this may decrease response rates.

12. Maintain a similarity in style among like items and responses. A similarity or parallelism within and among questions, choices, text cues, and presentations should be maintained for common or similar attribute measures. For example, if we start out by having positive attributions to the right and

negative to the left, it is usually a good idea to keep them that way throughout the questionnaire.

Numeral indexes offer another example. Respondents who see the first few high numbers as "good" may make a mistake and check the wrong number in items where low numbers are "good" and high numbers are "bad." If the writer changes cues, the writer must make it very clear when the "signals have been switched" and be very consistent with the use of signal switching cues.

There are other exceptions to the general rule of maintaining similarity in response formats. If we ask people to recall an extensive and detailed set of information and then follow this up with a second request involving another extensive and detailed consideration of information that is similar, performance on the follow-up question will be degraded. This is called "forward interference."

"Backward interference" can also occur. For instance, in complex questions, if we asked a third question of a similar nature, as one might do in a bridging or overall assessment question, we would again have problems, because the third question requires the respondent to retrieve information requested by the first question as well as by the second. The similarity of material in the second question interferes with the respondent's memory of the considerations he or she used to answer the first question; since this material is needed to answer the third question, the third answer is compromised.

Another exception deals with capacity. It seems that there is a limit to how much we can ask about a subject at any one time. People's performance starts to degrade after retrieving information on 20 or 30 similar items, even though hundreds or thousands of items are stored in their memory. If exhaustive recall

is required, the question writer should break up the topic into several questions and space these questions out.

The corollary to question and appropriate response similarity also holds. That is, if the appropriate response is dissimilar to how the respondent has learned to answer, performance will be degraded. To get around this, the style of question presentation should be changed or other cues included to let the respondent know that the appropriate response is different from that which was previously learned. However, regardless of the presentation switch, the question stem, text, and format should be in agreement and consistent within the question.

Remembering Frequency and Time of Occurrence

To measure frequency or time of occurrence, question writers need to relate the information about an event or series of events to a date or a specific time period. Questions measure frequency in two ways. They ask how many times the event occurred in a referenced time period (such as March 1 to April 15). This is called the "frequency method." Or they can ask when the event occurred and how much time elapsed before the next event occurred. This is called the "interval method." Interval measures often provide higher estimates, particularly if the reference periods are short and the period over which the measure is generalized is long.

Recall will be more accurate if the reference period (that is, the period for which data are requested) is short and the time gap between the reference period and that of recall is also short. However, as a practical matter, writers usually want the time periods to be as long as possible and often to extend back for a long time so as to efficiently capture as many events as possible. They also like to be able to deal with long and variable, different, or nonstandardized gaps

between the reference period and that of recall. But these requirements are incompatible with memory performance. Hence, the time periods are usually chosen on the basis of a trade-off between data accuracy, data-gathering efficiency, and data representativeness. The extent to which we can stretch the reference period and reference period reporting gap depends on the extent to which the events are salient or repetitive.

Salient events have been defined as events that are unusual or have significant economic and social costs or benefits to an individual. Events that have continuing consequences, such as President Kennedy's assassination, have been likened to snapshots by means of which exact details of the moment are remembered. Hospitalization, marriage, and car purchases are other significant events for which people have a high level of recall.

Although highly salient topics are less likely to be forgotten, they tend to be remembered as having occurred more recently than they actually did (this is called "forward telescoping"). Conversely, events that are less salient will be thought of as having occurred less recently ("backward telescoping"). For questions about the frequency or timing of salient events, respondents should be asked to report on events that occurred during the last 3 months. Periods of up to 6 months or a year have also been used. These longer periods help minimize telescoping.

If telescoping becomes a problem, there is still another approach that will help reduce it. First, we can ask that people recall the time period prior to the reference time period. This will capture the telescoped event. Then we can ask for a recall of the reference time period. This is called "bounded recall." For events of intermediate saliency, about 1 month is an acceptable compromise. These time periods seem

to provide the best trade-off for balancing omissions caused by forgetting and errors caused by incorrectly remembering an event against an efficient and representative time period.

Getting frequency data for repetitive events poses another type of problem. To get reasonably accurate data, question writers should use time periods of a few days or a week. However, while these periods may be more accurate, they are too short to be representative. If we want representative data, the periods should be from 2 to 4 weeks. These are less accurate but they will provide data that are more representative. For many purposes, the accuracy will be good enough to get a general idea of a pattern of events.

Respondents appear to use a different recall strategy for longer reference periods. They use their generic memories. That is, they report what usually happens, not what actually happened. Paradoxically, this may be more representative of normal experience.

Finally, if the event is neither salient nor repetitive, the time period should be very short, one day or at best a few days. Even at that, the recall accuracy may not be very good, and other methods should be used.

In summary, questions need to be asked in ways that help the respondents access their memories most efficiently and accurately and in ways that reduce their memory error. They must consider the short-term memory bias introduced by position, emphasis, and complexity or by the simplicity or similarity of the preceding text or succeeding answers. The question writers must know how the choice of time references, the saliency and repetitive nature of events, and the level of detail requested affect the accuracy of reporting. And, finally, they should take into account the limitations of a

respondent's memory—that is, the types of events and time periods for which recall is usually very poor.

Chapter 9

Minimizing Respondent Bias

The previous chapter discussed the response inaccuracies that can occur when evaluators inadvertently ask biased questions. Bias can also occur in the responses to questions because of a respondent's style in answering, such as the tendency to agree regardless of the issue, or because respondents perceive the questions as personally intrusive, objectionable, or threatening. This chapter discusses question writing techniques that help reduce or avoid these response distortions.

Response Styles

Response styles, or biases, have been defined as the tendency to respond in certain ways regardless of a question's content. Response styles vary considerably with the behavior in question and the conditions. For instance, respondents are more likely to answer questions about their education than their income. They are more likely to underreport problems about work while they are at work than while they are at home. They are likely to underreport behavior that is socially undesirable, especially if the behavior is presented in the extreme. Hence, question writers must be aware of response-style distortions and the ways to account for or counterbalance them.

Conversely, respondents may select socially desirable answers over other choices. Socially desirable responses represent culturally accepted norms for opinions and behavior.

Many people give socially acceptable answers about library card ownership, reading habits, charitable giving, and voting behavior. Occupation questions frequently provide another opportunity; occupational checklists with little or no explanatory details invite overstatement. For example, shipping clerks may check the job category "traffic manager," a position that can imply substantially more responsibility. Here

are two ways to reduce overreporting or overstatement of socially desirable responses.

1. The question writer should ask specific questions. For example, the shipping clerk will be reluctant to check "traffic manager" if answers to detailed questions about job responsibilities are required.

2. The question writer should make a single question containing a socially desirable response into two or more items. Respondents are more likely to answer truthfully about verifiable behavior. A series of questions can provide a respondent a "face-saving" escape. Although the behavioral question may not permit the respondent to give the most socially desirable response, topic awareness, knowledge, and other items may.

An example from a GAO audit illustrates these approaches. The Food and Drug Administration requires chemical testing and inspection. Asking chemists if they can do chemical tests could yield overreporting that they can. We did ask this question but, to assess the extent of overreporting, another question measured how much preparation they would need to do the tests. See figure 9.1.

Figure 9.1: Question to Reduce Overreporting

1. How much prior preparation would you require before performing tasks covered under compliance programs 7332.03, 7332.04, 7332.05, 7332.07, and 7332.10? *(Check one.)*

   1. ☐ No preparation required
   2. ☐ A brief check of the compliance programs
   3. ☐ One or two complete readings of the compliance programs
   4. ☐ A thorough review of the compliance programs with perhaps brief supplementary readings or consultations
   5. ☐ An extensive study of the compliance programs with detailed referencing and consultation

By taking the two questions together and interpreting the responses to both questions ("Can you do these tasks?" and "How much preparation do you need to do them?"), we could estimate overstatements of socially desirable alternatives.

Making a Good Impression

Respondents like to make a good impression. A study on personal bankruptcy illustrates the point. Individuals were asked to rate a list of factors on the extent that each contributed to their financial problems. The response "took on too many debts at one time" was underreported and "credit was too easy

to get" was overreported. To help overcome this tendency, do not place the sensitive items in prominent positions but list them midway in a checklist of several other plausible choices in a matter-of-fact manner. This approach can help respondents place response options in an objective frame of reference. Also, analyze the under- and overreported categories together. For example, "took on too many debts at one time" and "credit was too easy to get" were actually two sides of the same coin.

Extreme Points of View

Some people do not want to be categorized as holding an extreme point of view, even though they may feel strongly about an issue. When people are presented with three choices (for, neutral, and against, for example), they tend to select the middle category. To counteract this tendency, question writers can extend the scale to include more category ranges (definitely pro, more pro than con, neutral, more con than pro, definitely con).

However, some people select choices that represent extreme points of view regardless of the topic. Providing more category ranges (such as five or seven responses), organizing related topics so they are considered as a group, and providing adequate text to describe the categories (called "anchoring") help reduce a bias toward extremes.

Acquiescence

Because some respondents demonstrate the tendency to agree, writers should limit the use of agree or disagree questions. Besides offering the opportunity for a "yea saying" bias, they provide limited information. A more detailed discussion of those points and other problems associated with agree or disagree or Likert scales is presented in chapter 4.

Highly Sensitive Items

As mentioned in chapter 5, highly sensitive questions should be written with care and should be used only when the information is vital to the evaluation and cannot be otherwise obtained.

Personal questions, such as data on income, sex, marital status, education, and race, may be perceived by some respondents as intrusive and should be included only if necessary. Also, socially undesirable conditions, such as being unemployed or going bankrupt, may cause respondents discomfort. Other types of questions that can be perceived as threatening are usually highly specific to the topic under evaluation and the respondent's characteristics. Examples include surveying private industry officials about their bidding strategies, asking employees to assess the management of their agency or company, and asking self-evaluation questions such as "How would you rate your job performance compared with that of others?" Questions that could ask respondents to legally incriminate themselves should probably be reworded to remove this threat.

Before using sensitive items, the questionnaire writer needs to consider several questions: Can I get the answer I need through an archival source? How many people might not respond? Is the occurrence rate for the particular behavior or condition so low that asking for the data is not worthwhile? And how will the sensitive question affect GAO's image among respondents and the public? Having decided that sensitive items are necessary, the question writer should use the following guidelines to reduce underreporting and answer bias.

1. Explain to the respondent the reason for asking the question.

2. Make the response categories as broad as possible.

3. Word the question in a nonjudgmental style that avoids the appearance of censure or, if possible, make the behavior in question appear to be socially acceptable.

4. Present the request as matter of factly as possible.

5. Guarantee confidentiality or anonymity, if possible.

6. Make sure the respondent knows the information will not be used in a threatening way.

7. Explain how the information will be handled.

8. Avoid cross classification that would pinpoint the answers.

For example, when evaluators ask questions about income, respondents should be asked to choose from a list of income ranges rather than to enter specific dollar amounts. The income ranges should be appropriate for the target population and broad enough to afford the respondent a feeling of privacy. An example is in figure 9.2.

Figure 9.2: Question With List of Ranges

1. During the year in which you filed, approximately what was your gross annual family income from all sources? (That is, your family income before anything was deducted.) *(Check one.)*

   1. ☐ Less than $10,000
   2. ☐ From $10,000 to less than $15,000
   3. ☐ From $15,000 to less than $20,000
   4. ☐ From $20,000 to less than $25,000
   5. ☐ From $25,000 to less than $35,000
   6. ☐ From $35,000 to less than $45,000
   7. ☐ $45,000 or more

A series of questions and an indirect approach can diffuse the threat of asking about behavior that may be considered socially undesirable. For example, suppose that the evaluators need to find out about the job-hunting activities of the unemployed. The question series might be developed like the one in figure 9.3.

Figure 9.3: Series of Indirect Questions

1. Have you had any difficulty looking for work? *(Check one.)*

   1. ☐ Yes (Continue.)
   2. ☐ No (Go to 3.)

2. If yes, which of the following factors, if any, caused such difficulties that you were prevented from looking for work? *(Check all that apply or 8)*

   1. ☐ Illness (Go to 4.)
   2. ☐ Lack of transportation (Go to 4.)
   3. ☐ No child care arrangements (Go to 4.)
   4. ☐ No jobs available (Go to 4.)
   5. ☐ Age, sex or racial prejudice (Go to 4.)
   6. ☐ Employers won't give you a chance if you don't have years of education (Go to 4.)
   7. ☐ Other (Please specify.) (Go to 4.)

      _____

      _____

   8. ☐ Was able to look for work in spite of difficulties (Continue.)

3. Regardless of whether you had difficulties or not, how many contacts were you able to make?

   _____(number of job-related contacts)

Notice that items 1 and 2 recognize that looking for work is often difficult. This puts the respondent at ease, reduces the threat of revealing possibly embarrassing information, and makes not looking for work socially acceptable by providing very good reasons. This minimizes overreporting in looking for work.

Using a specified time reference can reduce a question's threat. For instance, if evaluators need to find out whether people are coming in late for work, the question writer can ask, "Were you more than a few minutes late for work this morning?" rather than "Are you usually late for work?" This is because people are more apt to admit to a single offense than to being habitual offenders.

The threat of some topics can be reduced if the rationale for asking the question is provided. For example, GAO wanted to send questionnaires to its disabled employees who needed the services of a federal program for the handicapped. The questionnaire's purpose was to assess the employees' work conditions and opportunities. The only way to identify people who needed handicapped-program services was to contact all employees who reported a disability to the agency when they were hired. However, many people consider a disability a private matter and might hesitate to answer the questionnaire. To encourage responses, we explained exactly why GAO management needed the information and how it would be used. Although we always state a survey's purpose, we explained this one more completely.

An example provided in chapter 8 illustrates another approach for potentially threatening questions. Private industry officials were asked to comment on competitive bidding strategies. To reduce the question's threat, we wrote the various bidding

strategies (conservative, realistic, and optimistic) carefully in a way that eliminated biasing terms such as "liar's dice." In addition, the question gave equal attention to all strategies, even though only one strategy was critical to the survey.

Still another way to reduce threat is to transfer or remove blame. For example, a questionnaire administered to a grief-stricken and guilt-ridden parent of a child with Reye's syndrome might ask, "Did your child take aspirin?" rather than "Did you give your child aspirin?"

Another technique to minimize overreporting of desirable behavior such as conducting compliance audits or voter participation is also to ask the respondent for concrete details associated with the desirable behavior. For example, we might ask, "If you conducted a compliance audit, please write the date, title of the audit, and name of principal auditor in the space provided. If not, check 'no audit conducted' and skip to 19." Similarly, we might ask voters to "List the address of the polling place" if they voted; if not, check "Have not voted" and "skip to 19."

There are also some ingenious ways of minimizing underreports of undesirable behavior such as failing to report all taxable income to the Internal Revenue Service. Two examples are the "randomized response technique" and the "list technique." Both methods use a prior or subsequently determined probability to mask the respondents' answer in such a way that the respondents can readily see that they are protected. Both methods still allow for population estimates.

Briefly, the methods work in the following ways. In applying the randomized response technique, we might ask only if the respondent's answers to the following two questions would both be the same or different: "Were you born in this month?" and "Did

you fail to report all your taxable income when you filed this year? Answer only the same or different." "(That is, if your answer would be NO to both questions or YES to both questions, check the same. Otherwise check different.)" Since we know the population probability of being born this month, we can calculate the population probability of not filing, but we cannot identify with certainty any individual who did not file.

As with the randomized response, the list technique uses a population conditional probability that can easily be determined to mask the certainty of the respondent's admission. For instance, the list technique might ask if any one of a number of infrequent events happened this month. For example, "Did you have a birthday, get a parking ticket, get a promotion, buy something that cost more than $300, underreport your taxable income, or take two or more plane trips this month? (Answer YES if any of these events happened last month)." Half the population would be asked the question with the illegal event included in the list and half without. The proportion of cheaters is calculated from the difference between the two populations.

However, a word of caution is in order. While these methods work and have been used to estimate behavior such as heroin usage, they have a down side. First, they are costly and more difficult to implement. Second, they need a larger sample size than other methods. Third, they should be undertaken only under the guidance of a skilled practitioner who is familiar with them.

Chapter 10

Measurement Error and Measurement Scales in Brief

The questionnaire is an instrument used to take measures. Virtually all instruments cause measurement error. For most physical measures, this measurement error is combined with all other types of error (for example, sampling error) to determine the total error. But questionnaire measures differ from physical measures because the instrument error and misspecification of variable errors are seldom determined. This is because such determinations take such laborious analysis and extensive testing and retesting that in many cases it is simply impractical to determine these errors. Hence, the convention is to report only sampling error and ignore the other sources of error, which in most cases are probably larger than the sampling error. The most practical way to address this problem is to use the guidelines presented in this transfer paper because they were specifically developed to minimize questionnaire measurement error. To make full disclosure, you should publish the questionnaire along with the sampling error so that report readers can get some idea of the quality of your measures.

Broadly speaking, there are two kinds of measurement error, bias and random error. Bias, sometimes called "systematic error" or "inaccuracy," occurs when respondents consistently underreport or overreport by a constant amount or range of amounts. For example, the phrasing of a questionnaire item about income may cause respondents to fail to include a particular category of income and consistently underreport. However, some surveys consistently overestimate. For instance, in some surveys the real level of unemployment is overstated because of the way the questions categorize people who are in transition between jobs.

The second kind of measurement error is called "random error" or, sometimes, "chance error," "unsystematic error," "noise," or "imprecision."

Respondents may react to a vaguely worded question in many different ways, some providing an answer that gives less than the true value and others an answer that is greater.

For example, we may want to know how many times a person visited a physician in the last year. If we asked, "How often have you sought health care?" our data would probably contain much random error. Some people might count visits to a podiatrist or a chiropractor and others might not. Some might count phone contacts, while others might count only office or hospital visits. Some might count a visit to a resort containing mineral springs. When a question is not precise about the information wanted, there is much opportunity for random error.

For mail questionnaires, every respondent reading the form should interpret each of the questions the same way. Also every question should be designed to minimize the biases that both the questionnaire and the questionnaire respondent place on the answers. This is why the preceding chapters emphasized the need for structure, the need for pretesting, and the need to consider the effects of format, appropriateness, qualifications, clarity, memory, and respondent bias.

Measurement Scales

In chapter 4, we discussed how different formats permit different levels of measurement. In selecting a question format, evaluators should think ahead to the point at which they will have finished data collection and will be starting their analysis. They should try to use the level of measurement or scale that will let

them use the preferred statistical techniques without prohibitively increasing costs or respondent burden.[1]

Equal-Appearing Intervals

Frequently, evaluators make observations on a variable for which the scale naturally has many small categories but they choose to use a coarser scale. For instance, people might be reluctant to tell their income (a fine scale), but they will tell if their income falls into a certain broad category. When using this technique, evaluators should try to make all the categories the same size. For example, the category "from $15,000 up to $20,000" is the same size as the category "from $20,000 up to $25,000"; both measure money in $5,000 increments.

Another example of the connection between the questionnaire format and the measurement scale can be seen in the Likert questions discussed in chapter 4. The Likert format has five broad categories. Should it be considered an interval scale, so that analysts can use the statistical techniques for interval data, or should it be considered only ordinal? The categories (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree) do not necessarily have equal intervals.

In the Likert format, we almost always treat the information as ordinal or ranking data. However, for some of the other intensity scales discussed in chapter 4, we can make a better case for an interval interpretation. For example, there may be some evidence that "generally satisfied" falls three quarters

[1]In chapter 4, we talked about categorizing, ranking, rating, equal interval, and ratio scales. In the first edition of this paper, we discussed these scales—called nominal, ordinal, interval, and ratio scales—in detail. But since then, GAO has published Quantitative Data Analysis: An Introduction, GAO/PEMD 10.1.11 (Washington, D.C.: June 1992), which deals extensively with this topic. Readers not well grounded in the use of these scales are referred to that document.

of the way between "very dissatisfied" and "very satisfied." However, even in such situations, it is usually best to show the proportion in each group and consider the category information as ranking data.

Since rating categories treated this way do not give much information, we sometimes make an additional effort to qualify the rating as quasi-interval data. When we do this, we call these categories "equal-appearing intervals," because, as best we can tell, the intervals appear to be equal. The equal-appearing interval formats use words, numbers, proportions, and behavioral anchors to make intervals that appear to be equal. For example, we could assume that "somewhat difficult" falls one fourth of the way between "no difficulty" and "extremely difficult."

However, such assumptions are very hard to justify. When making rating category scales, evaluators should be very careful to assign them on the basis of their knowledge of the variable in question, the literature, past experience, and pretest results. Sometimes it is a good idea to conduct a special study to verify assumptions. When uncertain about the assumption, evaluators usually treat the observations as ordinal data. If the assumptions are reasonable and the conditions are right, they sometimes treat attitude measures like "satisfaction" as interval data.

Chapter 11

Organizing the Line of Inquiry

As respondents begin their questionnaires, they discover the special language and the rules of the game, such as "skip to," "check one box for each row," and "if dissatisfied, go to question . . . ." This chapter suggests techniques for organizing a collection of questions into a well-designed instrument structured to elicit valid answers and to make the respondents' task easier. For example, several specific questions preceding a broad one can help respondents understand the range of factors to consider in making an overall judgment, and hard questions can elicit better responses if they are placed about a quarter or three quarters of the way through a long survey rather than at the beginning or the middle.

Setting Expectations

A set of instructions precedes the questions themselves. The instructions prepare respondents for the question-answering task in several ways:

1. They set a framework by identifying the data-gatherer, stating the purpose of the questionnaire, and describing the range and type of information needed.

2. They motivate respondents to answer by explaining the questionnaire's importance and relevance and protections of confidentiality or anonymity. (The pledge of confidentiality is discussed in chapter 14.)

3. They provide respondents in advance with some basic information, such as whether to designate answers by check marks or narrative responses, how long it usually takes to complete the form, and whether estimated or exact amounts are necessary.

Sequencing Questions

The instructions cause respondents to expect certain types of questions and the sequence of questions should fulfill these expectations.

Items should be presented in a sequence that is logical to the respondents. Frequently, the sequence mimics the flow of the process or condition under investigation. For example, in a study of printing industries, we would ask managers of firms for a description of a plant before asking for cost figures and ask for a description of equipment before inquiring about production data. If the natural or chronological flow of a topic is followed, the evaluators stand a better chance of helping respondents recognize and recall the information they need.

Using Subtitles as Cues

Related items that are grouped and accompanied by subtitles help the respondents quickly grasp the scope and nature of the inquiry. It also enhances the organizational flow and the cuing if the individual items within the group unfold meaningfully. For example, in a GAO evaluation on how personal bankruptcy cases were handled, we grouped the questions in accordance to the bankruptcy process and gave each grouping a subtitle. The first subheading was "bankruptcy proceedings." The question under this subheading followed logically "Under what name was the bankruptcy filed? Who filed the court papers?" and so on.

Choosing an Opening Question

The opening question should be interesting and highly salient to the topic, in order to capture the respondents' attention and demonstrate that their opinions are needed in key areas. It should introduce the language and rules of the questionnaire. Potentially objectionable and threatening questions should be avoided as initial questions. If possible, the

opening item should apply to all the respondents. Questions with such response options as "do not know" should be avoided. Respondents may feel uncomfortable about not being able to answer initial items or may question the relevance of the form to them.

However, in some instances, initial questions are used to determine whether respondents fit certain criteria and should complete the entire form. Respondents who do not meet the criteria should be thanked for their cooperation, told why their answers are not needed, and reminded to return their forms so that the population can be counted accurately. The following example illustrates how ineligible respondents might be notified:

"THIS SURVEY ASKS ONLY ABOUT CHILD CARE FOR CHILDREN UNDER 12. IF YOU DO NOT HAVE CHILDREN IN THIS AGE RANGE, DO NOT CONTINUE. THANK YOU VERY MUCH FOR YOUR HELP. PLEASE RETURN THIS QUESTIONNAIRE SO THAT WE CAN MAKE SURE WE ARE COUNTING YOUR RESPONSE IN OUR OVERALL POPULATION ESTIMATE."

A questionnaire should not be started with a broad or difficult question that will require a narrative response. Such questions require considerable effort to answer adequately. Also, the respondents have not yet learned enough about the information needed and may not provide the range and depth of data wanted.

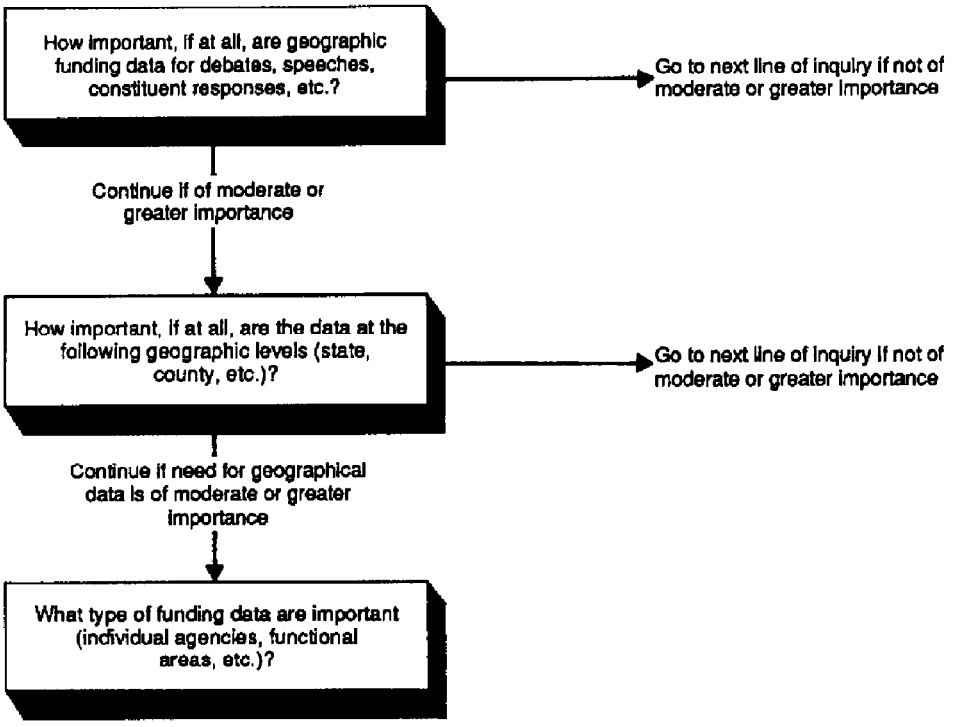
Sometimes trade-offs between question salience and ease of answering have to be made. In a survey of members of health maintenance organizations, a question asking individuals to rate their reasons for joining their plans would have been a natural starting point, but it could not be used as an opening question because of its complexity.

Demographic questions are usually placed near the end of a questionnaire if they may be perceived as highly personal and as perhaps less important to the questionnaire's purpose. However, if this is not the case, they may make a good starting question because they are easy. Also, the placement of demographic items depends on the topic and the audience. For example, military personnel are accustomed to providing rank and grade first. Also, if the demographic items seem less relevant to the questionnaire topic, the questionnaire designer may want to explain why this information is needed and how it will be used.

Obtaining Complex Data

Because a mail questionnaire is self-administered, it must be designed so that all or almost all respondents can faultlessly follow its instructions and feel that the form is easy to complete. For example, we surveyed congressional offices to measure their use of reports that show federal funding by geographic area. The reports provided information at various levels of detail (state, county, subcounty) and for a variety of data categories (individual programs, general functional areas, and so on). We needed to determine congressional use not only of geographical and funding categories but also of each particular combination (such as program data at the state level). In total, we needed 288 separate answers. Figure 11.1 shows how we broke down a complex question into individual items that would be easy to answer and that were sequenced logically. "Skip" and "continue" instructions accompanied each item and were set off in distinctive type to help respondents follow the item sequence.

Figure 11.1: Sequence of Questions Obtaining Complex Data



Using Transitional Phrases

Sometimes the respondent's task can be made easier by providing general information about the questions that will follow. Often, such text accompanies a subtitle and is used to alert the reader to a topic change. For example, in a survey of program managers of major weapon systems, a section of the questionnaire dealing with accountability was followed by a section dealing with the operating environment of acquisition personnel. Since this was

a topic change, a few lines of explanatory text were included to distinguish this section from the previous one.

Transitional phrases may be particularly necessary if a series of complex questions covers several pages. For example, in a survey of state coordinators for the mentally disabled, six pages were devoted to lengthy rating questions on the extent to which various federal programs encouraged or discouraged the deinstitutionalization of disabled populations. A few lines of text accompanied the section's subtitle, in order to explain the focus of the question series:

"FEDERAL PROGRAMS

Various federal programs provide institutional or community services to the mentally ill or mentally retarded. In the next series of questions, we ask you to tell us to what extent, if at all, various aspects of these programs currently encourage or discourage deinstitutionalization of the populations."

Warning respondents about a lengthy series of questions increases the number of items that will be responded to, because the respondents know each item will address a different program aspect.

Transitional phrases may also help respondents take a neutral point of view when making judgments. In a survey of an agency's employees in the field, respondents were asked to rate the benefits of rotation from a personal perspective and from the agency's perspective. To assist the respondents, transitional phrases were used. For example, after asking employees to rate rotation benefits from the agency's point of view, we wrote, "Now forget the office for a moment. How much do you think you would benefit personally from a rotational policy?"

Putting Specific Questions Before Overall Judgment Questions

Usually overall judgment questions seek to obtain an opinion that considers and weighs many factors. To get these considerations, it is best to precede a question with specific questions and items that deal with the facts, considerations, opinions, and experiences on which the overall judgment is usually based. For example, a question on overall job satisfaction should be presented after the respondents were asked to give separate consideration to the many factors that affect their job satisfaction: salary; benefits; job duties and responsibilities; supervisor, employee, peer, and subordinate relationships; career potential; and so on. A reversal of this order may prompt an unconsidered opinion to the initial response that colors the responses to the following, more specific queries. Subjects do not like to be inconsistent and will bias their subsequent responses to be consistent with their initial "top of the head response."

Put Filters Before Specific Questions

About half of the respondents who do not have an attitude, observation, experience, or knowledge about the topic will answer detailed questions as if they did. Answers from uninformed respondents cause error and may lead to false conclusions. One way to guard against this is to filter them out.

Before the line of questioning begins, the leading question might first ask if the respondents had an opinion and then ask the extent to which an issue was considered, giving operational definitions and anchors for a well-considered issue. Other alternatives might be to ask if the respondents had experience and the time and place of the experience or to ask about the respondents' qualification or role. These questions filter out and distinguish "no basis to judge" responses from uninformed answers.

The Influence of Context and Order

In organizing a questionnaire, the designers must realize that its context and order can affect responses to individual items. These reduce or increase response error, depending on whether they facilitate or interfere with the cognitive process. And while it is sometimes possible to assess the potential for such effects, in practice it is more difficult to predict their likelihood and whether these effects will help or hinder.

Contextual effects are usually seen in three ways: they influence the way the respondents interpret the question, the way they consider the tasks to be performed, and the way they may erroneously manipulate their answers to be consistent or different.

Context Affects Interpretation

Context cues influence the way the respondents make sense and meaning of individual questions, the way they interpret their recollection of previous questions as to what is and is not to be included in considering the answer of the current question. For example, in a survey on germ plasm, the contextual presentation assumed that the term "landraces" was always interpreted as traditional varieties; likewise, in a survey of businesses, "marginal" was always interpreted as "borderline" rather than as having something to do with the amount of money placed on deposit for purchase of stocks or unit costs after the production cost has been met. Also, as we have seen in the section above, putting specific questions before proceeding to the general helps the respondent recall and assemble the data needed for an informed judgment.

However, the tendency to respond to contextual cues can also have negative effects on the responses. For example, in one study we had several different questions that were introduced with qualifications and conditions under which the questions were to be

considered. Unfortunately, these qualifications were quite similar. The respondents, influenced by the similarity of the introductory qualification, thought all the questions were also the same. Also they failed to see the differences among each of the succeeding questions and did not answer. There are also some situations in which the question writer wants the respondent's "gut" reaction, first impulse, or unconsidered or unguarded response. The writer must then try to isolate the respondent from contextual cues such as the specific to the general organization that may interfere with initial and unconsidered answers.

Item interaction can occur even though various aspects of a topic are given equal attention. Inquiries that ask people to evaluate a topic from both a personal perspective and someone else's perspective might be difficult for respondents to answer neutrally but the order of the questions could help. As we mentioned earlier, when we surveyed an agency's field staff about rotation, we asked first about the benefits of rotation from the agency's point of view and only then about the benefits from the respondents' point of view, in order to obtain answers as objective as possible.

In some cases, interaction is associated with judgmental questions in which normative values play a role. In other cases, interaction may stem from how the scope of a general question is defined. Examples from the survey research literature can illustrate these points. If we were to ask respondents to report their degree of support for the rights of workers to strike and the rights of management to lock workers out, we would get different endorsement proportions, depending on how we sequenced the two questions. Endorsement for lockouts will be slightly higher if we ask first about a worker's right to strike. It is suspected that people use a norm of equal

treatment—if workers have a right to strike, business has a right to lock them out.

The effect of context cues of this type can also be difficult to predict. For instance, in two separate studies involving victims of crime, the factual questions about the crimes were asked before attitude questions because it was desirable to have the attitude reflect consideration of the specific crimes considered. Pretests showed that attitudes about crime did not change regardless of whether these attitude questions were asked before or after the crime-reporting incident questions. However contrary to expectations, the crime reports were more accurate if attitude questions were asked first. Apparently, the attitude response helped cue the memory search needed to question the victimization experience.

Context Affects Problem Solving

A second type of contextual influence affects respondents when solving the tasks asked by a questionnaire. The respondents, perhaps subconsciously, consider the order in which things are presented, how things are related, and the extent to which items receive greater or lesser emphasis according to the rules of conversational English. For instance, if the questionnaire has a number of questions about shop safety and then reintroduces the topic of safety in another part of the questionnaire, the respondents may not consider the subsequent questions as pertaining to shop safety. They will think these questions relate to safety in some other environment because, they will reason, if they did pertain to shop safety, they would have been presented with the earlier questions.

Respondents include or exclude according to their experience with conversational English. For instance if two qualifications or considerations are presented in serial order, the respondents will consider the

second to be most important. Two items to be considered in contrast should be paired with a conjunction like "but," not with words like "even so," which may not be seen as denoting contrast because such words are not usually part of conversational English. Two paired statements will often be seen as mutually exclusive. An example is "How is the morale of your work unit?" followed by the question "How is the morale of your organization?" In answering the latter question, many respondents will exclude their work unit from consideration when answering about the organization. Their rationale seems to be that we would not ask about their work unit separately if we intended it to be grouped with the organization.

People also group or differentiate things according to their experiences. They exclude items if they are not used to seeing them together, regardless of the logic of the specific requesters of the questions. For instance, they usually exclude the extremes of a classification. They also define constraints according to their own narrow range of experience. For example, to the citizens of Cumberland, Maryland, "Kelly," the local automobile tire company, was a big tire manufacturer. But to the citizens of Akron, Ohio, the home of Firestone and Goodyear, Kelly was a little company because they had never heard of it.

Context Affects What Is Included in and Excluded From Consideration

A third type of effect influenced by context and order is governed by the respondents' need to be consistent. To illustrate this, we can consider a previous example: "To what extent, if at all, are you satisfied or not with your job?" The question was followed by a series of 14 questions that asked about job satisfaction, taking into consideration such factors as salary, benefits, supervisor relationships, collegial relationships, and physical work environment. The first question is a broad and general question that might be answered off the top of the head. If this happens, respondents will bias all their successive

answers to the more detailed questions because they do not want to appear to be inconsistent with their first answer.

Conversely, respondents will sometimes erroneously exclude certain groups from consideration if the comparison contrast is very high, regardless of their need to be consistent. For example, in an attitude survey on driving, adult drivers excluded teenaged drivers when giving favorable assessments. In another study, a question on abortion on demand was less favored when preceded by a question on abortion in cases of rape and incest. This was presumably because the rationale for demand seemed more trivial when compared to the rationale for rape or incest. In a study of harassment in the military academies, male cadets may answer questions about their own harassment experiences quite differently from how they would answer if these questions were preceded by inquiries concerning harassment of female cadets. Moreover, some may exclude the females from consideration because they see them as very different from the rest of the body of cadets.

In designing a questionnaire, it is important to consider every part of the questionnaire where context and order can influence meaning, respondent problem solving, inclusion or exclusion, or contrast. After locating these areas of sensitivity, it is essential to conduct pretests. This is because the potential threats are not always realized and, when they are realized, the effect may not always go in the predicted direction.

Anticipating Respondents' Reactions

Except with very short forms, the attention, interest level, and effort of respondents fluctuate throughout the completion of a questionnaire. As respondents begin, they may be somewhat wary and uncertain. Specific expectations have been raised by the

transmittal letter and the instructions. Also, self-administered questionnaires resemble a test-taking situation in many respects. Respondents may wonder, "Can I follow the directions?" and "Where and how do I record my answers?" If the opening items are easy and nonthreatening, respondents become involved in the task and learn how to handle the format.

About one fourth to one third of the way through a form of average length, the respondent's interest and motivation are at high points. Complex items or questions that are critical to the survey can be introduced. Midway through the form, the respondent's attention and interest may waver. Less-demanding and less-critical items should be given at this point. Approximately three fourths of the way through the form, the respondent's effort and attention probably rise again. This accompanies a feeling that an investment has been made and what has been started should be completed. At this point, additional demanding and critical questions can be asked. Although this pattern of reaction may not always occur, it is applicable to many GAO forms, which tend to be moderately to very long.

Chapter 12

Following Quality Assurance Procedures

The quality of questionnaires can be checked by several methods, some of which are carried out during the design phase and others during the data collection or analysis phase. During the design phase, the questionnaire should be pretested on selected persons who represent the range of conditions likely to influence the evaluation's results. The questionnaire should also be sent out for review by experts who are familiar with both the issue area and the respondent group. Pretesting and expert review are some of the best ways to ensure that the instrument actually communicates what it was intended to communicate, that it is standardized and will be uniformly interpreted by the target population, and that it will be free of design flaws that could lead to inaccurate answers.

Validating, verifying, or corroborating responses; conducting reliability studies; and analyzing nonresponses are also important aspects of GAO's quality assurance effort. These tasks, which are conducted during data collection and analysis, are described in detail following pretesting and expert review.

Pretesting

By testing the questionnaire before it is distributed, evaluators can assess whether they are asking the right group of people the right questions in the right way and whether the respondents are willing and able to give the evaluators the information they need. Pretests are conducted with a small set of respondents from the population that will eventually be considered for the full-scale study. If respondents in a pretest have difficulty in responding or supplying information, it is likely that similar problems will arise in the full-scale study. If pretesting the questionnaire indicates that there is a low likelihood of obtaining accurate factual data sufficient for answering the assignment's objectives, troublesome questions

should be dropped or other techniques for data collection should be pursued.

Basically, pretests ask the following questions:

1. Is the content or subject matter of each question relevant to the respondent? Does the respondent have the experience and information to answer the question?

2. Are item-wording, phrasing, and other question construction components adequate to ensure that sound results will be obtained? Does the respondent understand the information request as it was intended? Are the response choices appropriate and comprehensive? Should the question be more specific? Is the time period suitable? Do filter questions and skip instructions work as planned? Are the instructions clear? Are transitions between sections smooth? How difficult is the questionnaire for the respondent? How long does it take the respondent to complete an item and to complete the entire questionnaire?

3. Are the questions asked in a way that will yield the needed information? Has a critical construct or variable been overlooked? Is the variable measured in sufficient detail?

4. Can and will the respondent give the evaluators the data they need? Can the respondent remember the type of information asked for in sufficient detail? If records must be consulted, how easily available are they? Is a question sensitive, objectionable, or threatening such that honest answers may be a major embarrassment or lead to possible punishment? Does the questionnaire adequately motivate the respondent to provide information?

Mail questionnaires are pretested by means of personal interviews. During the interviews, a wealth of information can be obtained by observing respondents as they complete the form and by debriefing them about the question-answering experience.

Who Should Conduct the Pretest?

In principle, the pretest should be conducted by a single person knowledgeable about both the pretest procedures and the questionnaire's content, because respondents are more apt to confide in a single individual than in a group. When this is not desirable, both an evaluator and a measurement specialist should be present. The evaluator addresses problems related to question content, and the measurement specialist assesses the questionnaire's overall adequacy as a data collection tool. Usually, the measurement specialist conducts the initial pretest while training the evaluator in observational and debriefing techniques. Such training is essential. After participating in a few sessions, the evaluator may be able to conduct the remaining pretests alone.

How Are Pretest Interviewees Selected and Contacted?

Pretest interviewees should be drawn from the universe being considered for the final study. The interviewees selected for pretesting should represent each of the major subgroups, conditions, and geographical or other units under investigation. The relevance and appropriateness of the questions may differ among these groups. For example, a national study of issues related to poverty should pretest the various groups of the poor in the universe—the elderly who are poor because of sickness, the elderly who are poor because they lack savings, the student poor, the disabled poor, and the welfare poor. Being poor in Maine may be quite different from being poor in Florida, so interviewees should be selected from

both states. Pretest subjects need not always be selected randomly.

A few people who are not typical of the universe should be interviewed in order to ensure the appropriateness of items for all potential respondents. For example, if the evaluators need to assess child-care arrangements made by employees, it is probably a good idea to test both extremes—a very large family and a family with only one child. Also, to test the questionnaire's readability, some interviewees should be selected whose language skills are somewhat less strong than those of the majority of potential respondents.

In principle, enough people should be tested to obtain a statistically valid sample of participants. However, time and staff resources are usually the controlling factors. For the typical questionnaire, between 8 and 12 pretests should be planned. This is merely a guide; sometimes we have had to manage with as few as 6 and at other times we have needed as many as 50. Exploring the particular needs of the survey with a measurement specialist helps determine the number of pretests.

The interviewees should be selected because they represent or have knowledge of the range of characteristics or conditions likely to be encountered—young and old, experienced and inexperienced, large and small companies, efficient and inefficient organizations, and so on. For example, in order to catch the range of conditions of the different streams of migrant workers as they moved northward, we pretested at the geographical beginning of the northward migration in Florida, Texas, and southern California and also at the middle and northernmost points.

If possible, the pretest subjects should be contacted by phone or letter and asked to voluntarily participate. They should be told what the evaluation is about, why pretesting is necessary, what the process consists of, and how long the testing is going to take. However, since they must do the pretest from the respondent orientation of a cold reader, they should not be given the pretest questionnaire in advance. Arrangements should be made to meet with each interviewee at a location as free from distraction as possible and at a time and place convenient for the interviewee. Of course, it sometimes happens that the pretest subjects cannot be contacted by phone. This would probably be the same with migrant workers or people coming through a customs border. In situations like this, volunteers must be recruited on site.

Care has to be taken in how a request for pretesting is communicated, because some people react with discomfort to the word "test." This kind of reaction can be allayed if the evaluators explain that the interviewee's comments and criticism are needed to test the questionnaire, not the interviewee. The lack of anonymity in a personal interview may also make the pretest candidate hesitant to participate. The candidate should be told that the information that will be provided will be treated confidentially and will not be included in actual data collection; evaluators are interested only in finding out how well the questionnaire works.

How Is the Pretest Conducted?

Pretesting has three stages: introductory comments, actual completion of the form by the interviewee, and debriefing.

Introductory Comments

The following points should be mentioned in the telephone contact and covered briefly again at the

beginning of the pretest session. The evaluators should

- state the role of the data collector;
- state the role of the person administering the pretest;
- state the purpose of the evaluation and the questionnaire and discuss the population to whom it will be sent;
- indicate the importance of the evaluation and the value of the interviewee's help in perfecting the questionnaire;
- remind the interviewee that responses are confidential;
- explain that pretesting involves the interviewee's completion of the form and will be followed by a short debriefing session to review the interviewee's comments, suggestions, and criticisms, explaining also that the interviewee will be given the same materials that would be received by mail, including a transmittal letter and the questionnaire form;
- state that the questionnaire should be completed as if it had been received by mail and no one else were present and mention that instructions on the form explain how to complete it and that the interviewee who cannot proceed without further explanation should stop and ask for assistance (interviewees should be encouraged to note on the form any problems or ideas that arise as the questionnaire is being completed);
- provide some examples of the type of item flaws or other problems the evaluators want the interviewee to look for (for example, an item may ask for dollar amounts by calendar year when amounts are available only for the fiscal year, or an item may ask for figures on the number of patients who were deinstitutionalized during a specific year but the institution's figures may count all the times each patient left who also entered more than once during the year, or the list of options may fail to include a critical component, or the interviewee may not be

sure of a particular response but no category such as "Not sure" has been provided for this, or a skip in the instructions may be confusing);

- tell the interviewee that the evaluators will be following the sequence of questions on their own copy of the form in order to monitor the flow of questions, thus addressing any potential concern the interviewee may have upon noticing that the evaluators are entering information on their form;
- state that frank and honest answers are appreciated and thank the interviewee for assistance; and
- conduct the pretests as a one-on-one session and not in a group. If the respondents suggest a group session, explain why this is not a good idea: in reality, the questionnaire would be read by a single individual working alone, the group interaction will influence everyone's understanding of the questions, some respondents are less likely to confide, and finally the pretester cannot handle more than one respondent at a time. There are exceptions, but we will discuss them in a subsequent part of this paper.

Completing the Questionnaire

The pretest administrator should carry out six tasks while the interviewee completes the form.

1. Recording the time it takes to complete each item. At the beginning of the pretest, the evaluators should position themselves so they have a clear view of the interviewee's questionnaire and face and as much of the body as possible. The start time should be recorded at the top of the evaluators' form. As the interviewee works, the evaluators should count silently the number of seconds it takes the interviewee to read the instructions or complete a question, and this time should be recorded next to the relevant section on their copy of the form. Evaluators should try to be unobtrusive. If the interviewee asks a question or the test is otherwise interrupted, the time taken out for the relevant item should be noted. Timing is obtained for two reasons: first, the average

time it takes all interviewees to complete an item serves as an index to the difficulty of items and, second, the average time it takes to complete the entire questionnaire serves as an index of respondent effort or burden.

2. Talking through. Some respondents feel comfortable talking out loud while answering. When they do, they should be allowed to do so and their verbalizations should be noted. They should be asked to say what is going through their minds while answering. However, many feel uncomfortable with talking through. Hence, this approach should not be used unless it feels natural.

3. Recording questions asked and clarifications made. When the interviewee asks a question, the evaluators should record key words or verbatim text as well as their own response next to the relevant item. These comments are used as an aid in debriefing and in item rewriting. Interviewees who are confused about what a question means should be provided a straightforward answer. Probing should be done during debriefing rather than during the test to see what the problem was. Evaluators should pay particular attention to how they answer any questions the interviewee raises, and they should be careful when providing explanations or alternative wording. In deviating from the prescribed text, evaluators may rephrase questions and bias the interviewee toward a particular response. However, if the interviewee is insistent and comfortable with discussing each question in turn and is giving good observations, the dialogue should be allowed to flow. Some people may recall with better insight if they speak as they see.

4. Noting nonverbal behavior. Evaluators should record any nonverbal behavior and body language that coincide with particular questions. Such behavior as hesitance in responding, facial expressions,

rereading questions, turning pages, and nervous movements (foot-tapping, fidgeting, and the like) may indicate item-design faults, question difficulty, or lack of relevance. Nonverbal observations are very important because they can be used as signals for questions that should be asked during debriefing. Methods for taking these observations will be discussed more extensively in the section on debriefing.

5. Noting whether instructions and format were easy to follow. Question instructions and format vary from item to item. Evaluators should notice how smoothly and quickly the interviewee reads directions and moves from one item to another. Did the interviewee ask questions about the instructions or the directions for filter questions? Could the interviewee follow the "skip to" or "go to" instructions with ease?

6. Noting erasures, uncompleted items, errors, and inconsistencies. These types of responses may indicate questionnaire design flaws. Evaluators can pick these up as they review the interviewee's questionnaire before debriefing.

Debriefing

The purpose of debriefing is not only to identify items that are difficult or misunderstood but also to get at the cause of these problems. The interviewee's answers and the evaluator's observations help uncover these problems and correct them. The debriefing usually takes 1-1/2 times as long as it takes to complete the questionnaire.

A debriefing should begin with a statement of its purpose, telling the interviewee that evaluators will be drawing on the interviewee's experiences and judgments to

- ensure that the intent of each item is clearly conveyed,

- evaluate the relevancy of items, and
- identify item-design deficiencies.

The interviewee's questionnaires should be reviewed in detail, and feedback to the evaluators' probing should be obtained. The major problems to look for are

- improper question format,
- inappropriate questions,
- improperly qualified questions,
- inappropriate language,
- failure to present an inclusive range of mutually exclusive alternatives,
- complex questions,
- unclear questions,
- question bias, and
- improper scales.

In discussing questionnaire items, GAO evaluators usually use the following sequence: (1) uncompleted items; (2) obvious errors and inconsistencies; (3) erasures; (4) items that took a long time to answer or appeared to cause difficulty; (5) items that took an unexpectedly short time to answer, possibly indicating that the interviewee missed certain key considerations; (6) questions the interviewee says caused uncertainty, undue deliberation, or difficulty; and (7) all other items not yet discussed. Alternatively, the sequence within the questionnaire may be followed.

The evaluators' approach in debriefing should be nondirective. They should try to elicit the interviewee's comments, problems, and reactions to the questionnaire without leading. They should use general comments to get the interviewee to reconstruct the questionnaire experience. For example, the interviewee's answers or the evaluators' observations of behavior can be used as a take-off

point: "You didn't answer . . . ," "You took a long time . . . ," "I noticed you seemed puzzled . . . ," or "Tell me what you had in mind when . . . ." Then the interviewee should be allowed to tell the reasons behind the behavior. Some areas may need a more direct approach. If "don't know" is the answer supplied, evaluators can probe to see whether the interviewee is being evasive. If evaluators believe the interviewee has an answer, they can push a little but not so much that a true "don't know" becomes a bad response.

During the observation period and debriefing, the evaluators should be very observant of the interviewee's paralanguage—that is, the vocal and facial expression, gestures, and body language used to modify speech. The evaluators should be careful about their own paralanguage so as not to send out conflicting messages or to send a message by this medium that may reinforce, encourage, extinguish, or inhibit the interviewee's comments. Instead, the message sent through an open and attentive posture, interested and pleasant facial expressions, soft, encouraging, motivating, and responsive voice should be that of a responsive person very interested in what the interviewee has to say.

The evaluators' posture should be open, facing the interviewee with a slight forward lean and attentive demeanor. They should sit to the side of the interviewee if possible. This signals team work and cooperation rather than competition while allowing the evaluators to see the interviewee's whole body. Before starting, the evaluators should assess the way the interviewee has arranged his or her space, and they should try to position themselves in accordance with the setting, avoiding an invasion of the interviewee's space. The evaluators should be seated at a comfortable conversational distance of 2-1/2 to 3

feet but close enough to observe the interviewee as he or she completes the questionnaire.

Eye movement is very important because that is the primary way of controlling the debriefing. When people converse, they tell each other when they want to talk or not talk and when they want others to talk or not talk by eye movement. Evaluators must learn to use this language. For example, looking at an interviewee is a signal that you want him or her to talk or keep talking. Looking away, occasionally offering slight gestures, stutter-stop interruptions, and throat clearing are signals to stop. A trailing voice, long pauses, silence and a head nod, and increased eye contact tell the listener that the speaker is finished with the topic and does not want to talk anymore. An increased rate, louder voice, filled pauses, halting gestures, and reduced eye contact tell the listener that the speaker is not finished and wants to keep talking. Skillful use of these signs will allow the evaluators to manage the interview, to get the interviewee to talk, and to avoid pushing him or her beyond his or her own knowledge limit.

Interviewees use paralanguage cues knowingly or unknowingly to tell interviewers what they think of the questionnaire, and these cues can often serve as a basis for conducting further probes. Some examples are speaking with variety in pitch and intensity, making pauses shorter than usual, speaking at an increased rate, and opening eyes wide to indicate involvement with the topic or certainty of the message. Hesitating speech, narrow eyes, longer pauses, many pauses, shrugs, slight side-to-side or up-and-down palm-down hand gestures, raised fingers or palms, furrowed brow, and a half smile or frown may indicate uncertainty about the information interviewees are providing or about their comprehension of a question. Slowly spoken, carefully enunciated, low-pitched speech with

hesitations can signal caution. Carefully enunciated speech without hesitation but with increased rate and intensity, narrow eyes, unraised eyelids, and knitted eyebrows may be signs of annoyance, difficulty, or dissatisfaction. Narrow pupils, raised eyelids, raised eyebrows, and frowns may show unpleasant surprise, while wide pupils and a smile can denote pleasant surprise. Extended looks or gazes to the side or at the ceiling often indicate that interviewees are thinking through the information, but if the gazes turn to blank stares, their interest has been lost or they are bored. Vague answers, shrugs, don't know signs, reduced head nods, reduced eye contact, and nervous twitches in the hands or feet may indicate deception. Looking at the interviewer's forehead may indicate a question. An increase in intensity and rate interrupted by an unexpected pause often signals that the next thing the interviewee says is very important. It is very important for interviewers to be aware of and observant of this paralanguage because perhaps as much as half of the communication that takes place between an interviewer and an interviewee uses this medium.

These observations generally apply to American culture and sometimes do not apply to other cultures or to individuals with certain disabilities. For example, in some non-American cultures, looking at an important speaker face-to-face is a sign of disrespect. There are also paralanguage variations particular to each of the ethnic American groups. Some have closer or more distant conversation spaces or look at their conversational partners somewhat more or less frequently than we described. However, these are differences in degree, and these guidelines will work for most situations for most American ethnic cultures.

Observing the interviewee completing the questionnaire with no direct queries before the

debriefing has the advantage of allowing for the assessment of contextual cues in a more realistic situation than other methods and for the use of aided recall without leading or biasing, but this method does have certain disadvantages. The interviewee sometimes loses spontaneity or forgets initial observations or first impressions. Also interviewees who cannot articulate a rational explanation for a feeling or perception may make one up. Therefore, if the interviewees really want to talk about each question as they read, evaluators should permit them to do so. It is also helpful to do at least one or two pretests using a "talking through" approach.

When the debriefing has been completed, interviewees should be thanked for helping to perfect the questionnaire. As soon as possible, the evaluators' comments and observations about the pretest should be recorded.

Standardized Pretests

Except for the flexibility granted to the interviewer to probe the subject according to his or her pretest observations, the pretest protocol should be standardized as much as possible. The main reason for standardization is to promote a sufficient number of replications to evaluate the pretest findings. However, in certain circumstances, it may be more efficient to revise the pretest instrument in midstream. This usually happens in the following situations: (1) when the design errors are so obvious that there is little doubt about how to make the correction; (2) when the initial instrument is far off the mark (an example is when more than one third of the questions need major revision); (3) when the corrections are so difficult to make that the question writer is not certain as to whether he or she has fixed the problem. When these situations occur, it is better to revise the instrument and begin a second round of pretests.

Expert Review

Because GAO's studies are wide ranging, we frequently need to seek outside comments on the questionnaire approach. The purpose of this expert review is twofold. First, we want to determine whether the questions and the manner in which we ask them are adequate for addressing the larger questions posed by the evaluation. Second, we want to find out whether the target population for the survey has the knowledge to answer the questions. In many instances, the agency officials whose program is under review can help provide this information.

People who provide expert reviews do not act as pretest interviewees; they do not answer the questions but provide a critique. Only on rare occasions does a reviewer serve as a pretest subject, too. The expert must have a thorough knowledge of the target population. For example, in a study of the Foreign Corrupt Practices Act, a former head of the Securities and Exchange Commission served as an expert. In a survey on indirect costs of research grants, we sought the help of the president of the National Association of College Business Officers, because most research grants are administered by members of this society.

Validation and Verification

Validation is an effort to ensure that the questionnaire is actually measuring the variables it was designed to measure. Validation is important because if the questions are not valid measures of the constructs we are studying, even answers verified as accurate will not provide us with the quality data needed for our findings, conclusions, and recommendations.

Verification is a way of checking or testing questionnaire answers with records or direct observation to reduce the risk of using data that are inaccurate. Verification is different from validation. For example, suppose we are interested in the quality

of health care and propose the number of visits to a doctor as an indicator. To validate, we would have to show that the number of visits could be taken as a measure of the quality of health care. And in proving this, we are likely to find that this indicator is valid only under certain conditions. However, if we wished to check the accuracy of the patients' self-reports as an estimate of the numbers of doctor visits, we might compare this estimate with physicians' records. In doing this comparison, we are testing the soundness of self-reports only as a measure of visits (verification), not as a measure of the quality of service (validation). Verification tells us if the subjects self reports can be trusted as an accurate measure but not necessarily as a valid measure. Verification is ideally conducted by testing a population sample. Since this is not always practical, GAO often shows that other comparable studies had similar findings or cross-checks for internal consistency.

Corroboration (referred to as validation in some circumstances) of questionnaire results against similar information from another, independent source can also provide supporting evidence to increase confidence in the relative accuracy of questionnaire data.

The reliability of questionnaire results tests whether a question always gets the same results when repeated under similar conditions. Answers can be highly reliable without being either verified or valid.

Why do evaluators have to validate, verify, corroborate, and make reliability checks? GAO has to do much of this work because most of the time it cannot use "standardized" instruments—those that have already been tested during their development. We are either measuring things that have not been measured before or measuring previously measured

things under different circumstances. Since we most often do our own instrument development work, these essential attributes are discussed in more detail.

Validation of the Questionnaire

To validate we show that the observation measures what it is supposed to measure. The best way to demonstrate validity is to demonstrate the relationship between the measurement and the construct being measured in a setting as controlled as possible. This is called "construct validation." For example, we wanted to use the time it took to complete questionnaire items as a measure for the construct "item difficulty." To validate this, we deliberately constructed sets of items that varied in difficulty by changing the reading levels, the concepts, the memory requirements, the decisions, and the operations until we had developed a set of items that spanned the range from easy to extremely difficult. Then we administered this test to a number of people under controlled conditions. We measured the time to complete the item, the number of mistakes (another possible measure of difficulty), and the respondent's ratings of the difficulty of the items. As the difficulty of the items increased, so did the mistakes, the respondent's ratings of difficulty, and the response times. We concluded that the time it took to complete an item could be taken as a valid measure of the item's difficulty.

In another study, evaluators used supervisory ratings as a measure of employee performance. To validate this, the evaluators compared the supervisors' ratings of employees with employee performance test scores. These performance tests were conducted independently of the supervisory rating.

Few measures are completely valid; the more rigorous and varied the validity tests are, the stronger is the case that can be made for a measure. There are

a number of other ways to test validity. Although most of them are less convincing than construct validation, they are easier to apply. But no validity assessment is perfect, and no single method is best suited for all situations.

A very practical method of assessing validity is to use "content validity." In this approach, evaluators might ask experts to make sure that the measure includes the content they want to measure. For example, in a study of the Financial Integrity Act, several measures of financial integrity were proposed: time since audit, number of audits, amount of cash, cash controls, ease of access to cash, number of people with access to cash, and so on. Financial accounting experts reviewed the measures and concluded that they would be valid indicators of financial integrity.

Prediction is also used to assess validity. For example, in one study, we developed an instrument that would measure the restrictiveness of zoning laws and practices. We validated the measure, in part, by showing that the restrictiveness score was correlated with land-use patterns.

Criterion comparisons are also used. For example, if a new test is supposed to measure intelligence, then the people who take it ought to get similar scores on the Stanford-Binet IQ test (a time-honored and extensively validated test).

Validity can be tested by looking at the relationships between factors that should be positively correlated or negatively correlated. For example, measures of the quality of training ought to correlate positively with productivity. If they do, we have some confidence in the validity of the measures. The measure of a participative management style ought to correlate inversely with a measure of an authoritative

management style. If it does, confidence in the validity of the measure is strengthened.

Although the rigor and pluralism of methods that are used determine the credibility of a claim for validity, resources are often limited. We tend to validate most often when the measures are complicated and abstract, or unproven, or critical to the study findings and likely to be challenged.

Verification

Our measures must provide accurate data. We test for this precision by comparing the data against an accurate source, by putting in controls that reduce observation errors, or by repeating the measurement process. This practice is often called verification or corroboration.

Determining how much verification should be done to ensure the quality of data obtained through questionnaires is a management judgment. The extent of verification should be based on the type of data, its use as evidence to address the assignment's objectives, the relative risk of it being erroneous, and alternatives available to verify data, including time and resource constraints.

Opinions and attitudinal data, on the one hand, are testimonial evidence and could ideally be verified by checking the consistency of the answers with actual experiences and behavior. However, this is not often easily done and may not be necessary since the data are presented only as opinion. Factual data, on the other hand, can be verified through observation, cross checked with other witnesses, or checked against records.

The most convincing method of verification is to compare on a test basis the respondent's answers with evidence developed from an "on-site inspection"

that involves direct observation or a review of records. Such verifications are ideally conducted on a statistical sample of the respondent population. Practically, a judgment sample considered typical of the population is often used.

In addition to on-site verification, or when such verification is not practical, the following types of steps can be taken to raise the evaluators' level of confidence in the reasonable accuracy of the data:

- Ask respondents to send a copy of specific records when they return the completed questionnaire.
- Telephone and obtain clarification from respondents who provided important data that seemed out of line when compared to the data provided by similar respondents.
- Telephone a random sample of respondents and attempt to ascertain the extent to which they consulted appropriate records to obtain the most significant factual data provided in their responses.
- Corroborate or verify through other data bases, records, or prior reports.
- Corroborate the questionnaire results by comparing them to the results of similar studies or having them reviewed by outside experts knowledgeable about the program or topic.
- Cross check aggregate statistics from the questionnaire against data reported by other organizations.
- Include consistency checks in the questionnaire by asking for the same or similar information in more than one question.

Another aspect of verification is checking the accuracy of keyed data by comparing the keyed records with the original source records. Data entry operators verify by keying in the source document twice and check to see if they get identical answers each time. However, GAO sometimes also uses

controls to verify the accuracy of the data entry such as checking for illegal codes or out-of-range values.

Initial plans for verification should be part of the data collection and analysis plan that is completed during the design phase. The type and amount of verification should be appropriate for ensuring that the evaluators will have sound evidence to address the assignment's objectives. The initial plans may need to be modified when the questionnaire is pretested, when the questionnaires are returned and the responses are being analyzed, or whenever there is reason to doubt the accuracy of the questionnaire results.

Testing Reliability

"Reliability" refers to the consistency of measures. That is, a reliable measure is one that, used repeatedly in order to make observations, produces consistent results.

Testing reliability is difficult, and expensive, because the evaluators have to either replicate the data collection or return to those who were questioned before. People do not like to be retested. Because of this, GAO often does not test reliability if we have good reason to believe our measures are stable. If we cannot make this assumption with a high degree of certainty or if we are likely to be challenged on this issue, we should test this assumption.

Some situations in which the reliability testing of the questionnaire should be conducted follow. First, if the respondents as a group lack motivation or interest, they may not invest much care or thought in the questionnaire and their answers may vary randomly over time. Second, if respondents are expected to purposely exaggerate, retesting sometimes brings a more sober reconsideration. Third, for some topics, asking respondents to complete the questionnaire at home may produce different results from having them

fill it out in another setting. For example, a questionnaire on military reserve training completed at home produced different answers than one completed while reservists were at summer training with their units. Fourth, there is a tendency when most respondents take an extreme position for extreme values to drift toward the norm when the measures are repeated at different times.

It is important to note that the procedures for testing the reliability of answers are different from those for verifying answers. When information is verified, evaluators usually go to a different source for the same information or use a different technique on the same source, such as observations or in-depth interviews. To test reliability, evaluators have to administer the same test to the same source.

Analysis of Questionnaire Nonresponses

Item and questionnaire nonresponses also must be analyzed because high or disproportionate nonresponse rates can threaten the credibility and generalizability of the findings. Suppose only half the people respond. Nothing is known about the other half. In particular, the reason they did not respond may be related to important differences between them and the responding group. Usually, if the nonresponse rate is small, we can make plausible assumptions that discount the potential effect of the nonrespondents. However, even in this case we should use whatever prior information we have to check for systematic differences between respondents and nonrespondents.

In mail surveys, it is rare to get an answer from every questionnaire recipient. Some people in the sample may never have received the questionnaire. Some who did will choose not to answer. Thus, the original sample of recipients can be expected to shrink somewhat. The real problem is not so much the

decreased sample but whether those who chose not to answer had disproportionately different views from those who did. For example, most of those who did not respond might have been opposed to something favored by those who did. We would then mistakenly believe in the generalizability of our sample responses, unless we investigated the reasons for nonresponse. This would threaten the representativeness of the sample and the ability to generalize from the sample to the population.

In GAO, we account for all questionnaires mailed or interviews attempted in our workpapers and in our products. This includes the number of questionnaires returned or interviews completed, the number of intended respondents who refused, the number of questionnaires that were undelivered or interviews that could not be conducted, and so on. We calculate a response rate that is the percentage of eligible study cases drawn from the sample or population list that provide usable data. We also obtain and analyze information about all nonresponse groups to determine how they differ from those who did respond. (The current policy guidance on accounting for survey responses is included in chapter 10.5 of GAO's General Policies/Procedures Manual.)

In order to make plausible generalizations, the effective response rate should usually be at least 75 percent for each variable measure—a goal used by most practitioners. By effective response rate, we mean the percentage of people who return the questionnaire minus the percentage of people who failed to answer for the variable in question. Small to moderate differences between the respondent and nonrespondent populations will then usually have little or no bias effect on the results. Transmittal letters that convey the relevance and importance of the questionnaire and systematic follow-ups help bring high response rates.

Nevertheless, the nonrespondent population should be analyzed unless the response rate is over 95 percent. A comparison of respondents and nonrespondents with regard to demographic and other important characteristics can reveal whether or not nonresponse occurred systematically (for example, in a particular region or other segment of the questionnaire group). In a survey of employees who were subject to an agency's reduction in force, we found a high nonresponse rate in the Atlanta region. In another survey on block grants, all respondents whose last names began with "U" were missing. In both surveys, the mailgram contractor had neglected to send out follow-up notices. This could have resulted in misrepresentation of the respondents' views, insofar as the groups that were excluded differed from those that were included.

Aside from reflecting mailing mistakes, the nonresponse rate may reflect certain conditions or respondent attributes. In a study of zoning and group homes, we analyzed responses to see whether people from states with unfavorable zoning laws did not respond. We also compared response rates for the types of population that facilities served (for example, the mentally retarded or emotionally ill).

The workpapers should document the analysis of the composition of the nonrespondents, indicate the number and type of categories excluded from the expected population or sample, and document attempts to verify or trace the correct addresses of those who could not be reached by mail. If a nonresponse bias is detected, and we can make assumptions about the nonrespondent population, the survey results should be adjusted. For example, if a disproportionate number of nonrespondents are from California and we can assume that they are no different from the California respondents but we find that the people from California respond very

differently from people in the rest of the nation, we should weight the California responses to account for this underreporting bias.

If the response rate is lower than 75 percent and the standard follow-up procedures have been followed, it may be necessary to telephone or interview a random sample of nonrespondents to obtain answers to key questions or to find out why they did not complete the form. This information is important for two reasons: it brings more confidence to the evaluator about the meaningfulness and systematic nature of the nonresponses, and it helps assess the data that were returned. A discussion of the nonresponses should be included in the workpapers and in the discussion of methodology.

In addition to the people who do not return the questionnaire, some proportion of the people who do respond will not complete some items. Thus, the average nonresponse rate should also be calculated for each item in order to determine whether the data from an item can be included in the analyses.

Item nonresponse rates average about 3 percent. If the rate is more than about 7 percent, it should be analyzed to determine if the item presented a threat to respondents, was not perceived as relevant to the questionnaire focus, or contained design flaws or other factors that caused the low response rate. If the nonresponse rate is uncharacteristically large and, consequently, the item is excluded from our analysis, the final report should disclose this. Again, the item nonresponse analyses should be included in the workpapers and the discussion of methodology.

Chapter 13

Designing the Questionnaire Graphics and Layout

A questionnaire should be easy to read, attractive, and interesting. Good graphics design and layout can catch the respondent's attention, counteract negative impressions, cut the respondent's time in half, and reduce completion errors. If the design format works, respondents will feel they have received an important document outlining a reasonable request on which they should act.

The front page of a GAO questionnaire has a title, instructions, and logo or seal. The text of the instructions should have two columns to promote ease of reading. At the normal reading distance, the eye cannot span much more than 4 inches without refocusing, and most people cannot immediately take in more than seven to nine words in a single glance. A string of seven to nine words with the type size GAO usually uses (10-point type) usually takes up 3-1/2 inches. Furthermore, the two-column format gives the page a formal and patterned look.

To reduce bulk, both sides of a page are often printed. Usually, the pages are stapled in the upper left corner to look more like a letter and better suit the mail-out package. Booklets are used when a sturdier construction is needed or when the respondent has to refer back and forth to related questions. The questionnaire may or may not have a cover.

Instructions

The first part of the questionnaire should present the introduction and instructions. Because the transmittal letter is frequently separate from the questionnaire, instructions should repeat some of the material in the transmittal letter. The instructions should

- state the purpose of the survey;
- explain who the data collector is, the basis of its authority, and why it is conducting the survey;
- tell how and why the respondents were selected;

- explain why their answers are important;
- tell how to complete the form;
- provide mail-back instructions;
- list the person to call if help is needed to complete the form;
- provide assurances of confidentiality and anonymity when appropriate;
- tell how long it will typically take to complete the form;
- explain how the data will be used;
- explain who will have access to the information;
- disclose uses that may affect the respondents; and
- present the response efforts as a favor and thank the respondents for their cooperation.

The instructions should be concise, courteous, and businesslike.

Questionnaire Format Preparation

Most GAO questionnaires and most pretests are reproduced from texts prepared on word processors. Computer programs such as QUEST, Wordperfect, and other desk-top publishing packages convert this to typographic text suitable for publishing. Sometimes the text is typed directly in publishable form. These texts are almost as attractive and readable as texts prepared by commercial printers and they are quicker and cheaper to produce. However, an attractive, readable, and business-like style and type should be used. Documents that look official, professional, and inviting are likely to be answered. Good layout and composition can cut reading time in half and can reduce the respondent's burden. This is particularly important when

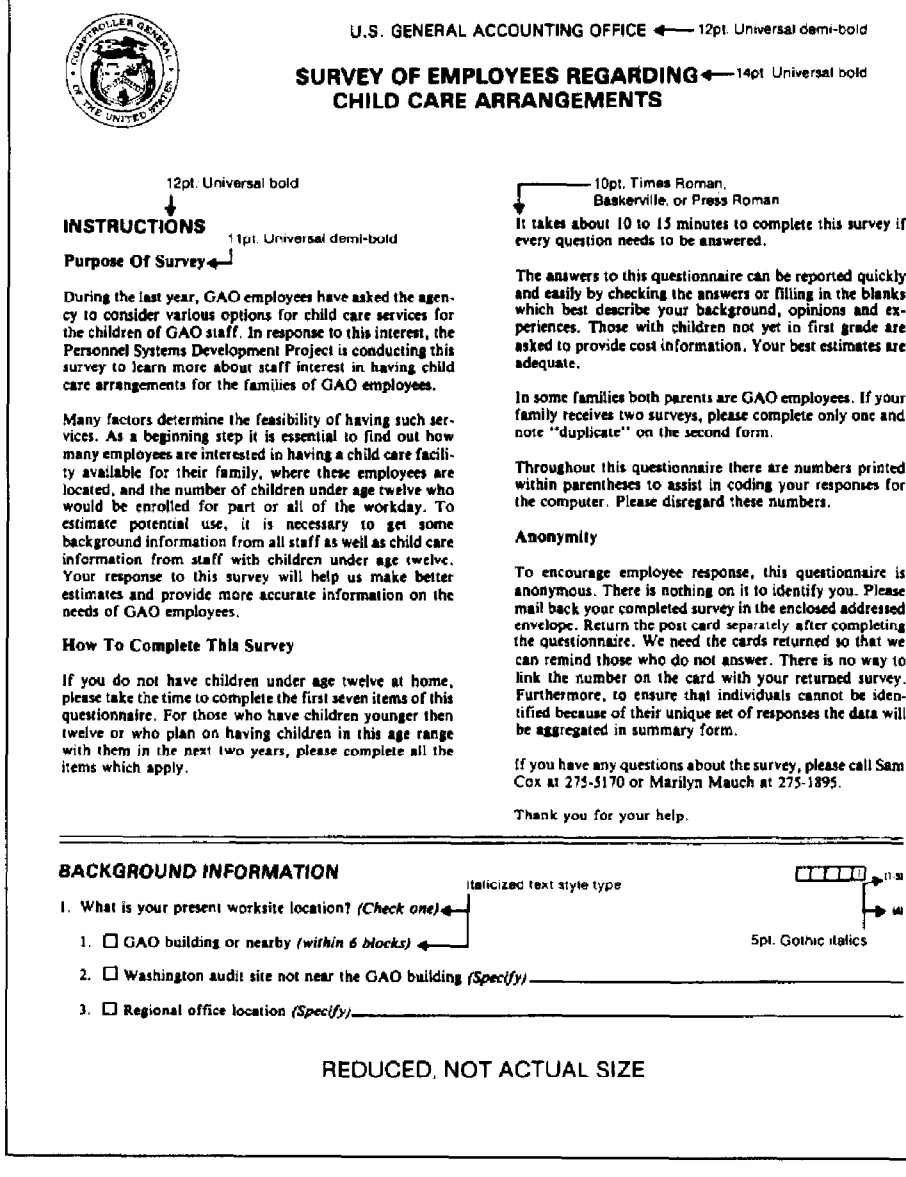
- the respondent group has low literacy,
- the questionnaire is very long and complex,
- a large population is being surveyed,
- a prestigious group is being addressed, or

- the data collector's professional image is very important.

Typographic Style

The size, style, and density of type are signposts to guide the respondent's eye and to signal the kind of information being presented. An example is in figure 13.1.

Figure 13.1: Partial Questionnaire



U.S. GENERAL ACCOUNTING OFFICE ← 12pt. Universal demi-bold

SURVEY OF EMPLOYEES REGARDING ← 14pt Universal bold
CHILD CARE ARRANGEMENTS

12pt. Universal bold

INSTRUCTIONS

11pt. Universal demi-bold

Purpose Of Survey

During the last year, GAO employees have asked the agency to consider various options for child care services for the children of GAO staff. In response to this interest, the Personnel Systems Development Project is conducting this survey to learn more about staff interest in having child care arrangements for the families of GAO employees.

Many factors determine the feasibility of having such services. As a beginning step it is essential to find out how many employees are interested in having a child care facility available for their family, where these employees are located, and the number of children under age twelve who would be enrolled for part or all of the workday. To estimate potential use, it is necessary to get some background information from all staff as well as child care information from staff with children under age twelve. Your response to this survey will help us make better estimates and provide more accurate information on the needs of GAO employees.

How To Complete This Survey

If you do not have children under age twelve at home, please take the time to complete the first seven items of this questionnaire. For those who have children younger then twelve or who plan on having children in this age range with them in the next two years, please complete all the items which apply.

10pt. Times Roman, Baskerville, or Press Roman

It takes about 10 to 15 minutes to complete this survey if every question needs to be answered.

The answers to this questionnaire can be reported quickly and easily by checking the answers or filling in the blanks which best describe your background, opinions and experiences. Those with children not yet in first grade are asked to provide cost information. Your best estimates are adequate.

In some families both parents are GAO employees. If your family receives two surveys, please complete only one and note "duplicate" on the second form.

Throughout this questionnaire there are numbers printed within parentheses to assist in coding your responses for the computer. Please disregard these numbers.

Anonymity

To encourage employee response, this questionnaire is anonymous. There is nothing on it to identify you. Please mail back your completed survey in the enclosed addressed envelope. Return the post card separately after completing the questionnaire. We need the cards returned so that we can remind those who do not answer. There is no way to link the number on the card with your returned survey. Furthermore, to ensure that individuals cannot be identified because of their unique set of responses the data will be aggregated in summary form.

If you have any questions about the survey, please call Sam Cox at 275-5170 or Marilyn Mauch at 275-1895.

Thank you for your help.

BACKGROUND INFORMATION

Italicized text style type

5pt. Gothic italics

1. What is your present worksite location? *(Check one)*
   1. ☐ GAO building or nearby *(within 6 blocks)*
   2. ☐ Washington audit site not near the GAO building *(Specify)* _____
   3. ☐ Regional office location *(Specify)* _____

REDUCED, NOT ACTUAL SIZE

Figure 13.1: Partial Questionnaire (Continued)

PRE-FIRST GRADE CHILDREN

14. Please list the age of each child living with you who has not entered first grade, and the types of child care presently provided during your workday. We also need time and cost information. Please list the usual number of hours of care per workday, the number of days of care per week, and the weekly cost.

List age of child. Use a different row for each child.	Type of Care Provided (Check all types of care usually provided for each child during the workday)						Child Care Provided (Include care by relatives. Report information to nearest $ or hour.)		
							Amount		Cost
	By spouse or relative at your home	By spouse or relative at their home	By sitter at your home	By sitter at sitter's home	By staff at child care center, nursery school, kindergarten or group home	Other (Specify)	Total Number of Hours of Care During Workday	Total Number of Days Per Week Child Care Provided	Total Cost for Care of Child Per Week (Include all related costs, except transportation)
	1	2	3	4	5	6			
1. (yrs) (mos)									$
2. (yrs) (mos)									$
3. (yrs) (mos)									$
4. (yrs) (mos)									$
5. (yrs) (mos)									$
6. (yrs) (mos)									$

10pt. Gothic

10pt Gothic bold

Shading

IF YOUR CHILDREN ARE CARED FOR BY A FAMILY MEMBER (E.G., SPOUSE, RELATIVE, ETC.) OR YOU DO NOT HAVE ANY CHILD CARE COSTS, GO TO QUESTION 16.

*About a dozen children cared for

Figure 13.1: Partial Questionnaire (Continued)

15. *Consider all types of child care services that you use. Overall, how satisfied or dissatisfied are you with the following features of the service(s)? (Check one column for each feature)*

Features of Child Care Service	Very satisfied	Generally satisfied	Marginally satisfied	Generally dissatisfied	Very dissatisfied	
	1	2	3	4	5	
1. Reliability of service (e.g., dependability, open according to schedule, etc.)						(14)
2. Hours and days service available or months of the year						(15)
3. Convenience of services (travel time and distance)						(16)
4. Safety and well-being of children						(17)
5. Quality of care, staff, program and facility, etc.						(18)
6. Cost of care						(19)

16. About how many miles is it to your worksite one-way? Also, about how long does it usually take you to get there? *(Exclude time needed to transport children for child care, if applicable.) (Complete both items.)*

_____ *(miles to worksite one-way)* (22-24)

_____ *(one-way trip time in minutes)* (25-27)

17. How much time does it usually take you or a friend or family member to transport your family one-way to child care services?

_____ *(Daily estimated time one-way in minutes)* (28-30)

18. How do you presently get to work? *(Check all that apply)* (31)

1. ☐ Bus
2. ☐ Subway
3. ☐ Carpool
4. ☐ Drive separately
5. ☐ Commuter Train
6. ☐ Other *(Specify)*

_____

_____

REDUCED, NOT ACTUAL SIZE

Figure 13.1: Partial Questionnaire (Continued)

19. GAO has been asked to consider child care services for employees. If a child care facility is available for your family in the next two years, how interested or not are you in using it? *(Check one)* (32)

1. ☐ Of no interest to me
2. ☐ Of little interest to me
3. ☐ Of some interest to me

(GO TO QUESTION 31)

4. ☐ Of moderate interest to me
5. ☐ Of great interest to me
6. ☐ Of very great interest to me

(CONTINUE)

20. Which type of location for child care do you prefer—a location at or near your worksite or a location near your home? *(Check one)* (33)

1. ☐ At or near worksite **(CONTINUE)**
2. ☐ Near home **(CONTINUE)**
3. ☐ Either a worksite or a home location is acceptable **(GO TO QUESTION 22)**

21. If you could not get the location you prefer *(the location checked in Question 20)*, are you still interested in enrolling your child (or children) at the other location? *(Check one)* (34)

1. ☐ Yes
2. ☐ No

22. Assume you were able to get the location you prefer. If a high quality service for pre-first grade children opened in the next two years, how many of the children in your care would you enroll on a regular basis? *(If none, enter "0" (zero) and go to Question 31.)*

_____ *(number of children)* (35-36)

23. How much would you be willing to pay weekly for a child to receive child care conducted for GAO families during working hours? *(If part-time, report for hours of care needed during week.)* (37-38)

1. ☐ Less than $30.00
2. ☐ From $30.00 to $34.00
3. ☐ From $35.00 to $39.00
4. ☐ From $40.00 to $44.00
5. ☐ From $45.00 to $49.00
6. ☐ From $50.00 to $54.00
7. ☐ From $55.00 to $59.00
8. ☐ From $60.00 to $64.00
9. ☐ From $65.00 to $69.00
10. ☐ From $70.00 to $74.00
11. ☐ $75.00 or more

24. Would you still be interested in using such child care services if fees were based on your family income? That is higher income staff would pay slightly more (e.g., 5 or 10% more) than the average cost per child and lower income staff would pay somewhat less. *(Check one)* (39)

1. ☐ Definitely yes
2. ☐ Probably yes
3. ☐ Uncertain
4. ☐ Probably no
5. ☐ Definitely no

REDUCED, NOT ACTUAL SIZE

The title is the most noticeable feature on the questionnaire's front page. It should be a short statement (12 words or less) that identifies the population from which information is sought and gives a clear idea of what the questionnaire is about. Because of its importance, it should be printed in large type (for example, 14 point in bold). GAO uses Universal, or a similar typeface, because it is official-looking and easy to read in bold capital letters. (Usually, capital letters are much more difficult to read than lowercase letters.)

Another feature of the title page in GAO questionnaires is GAO's logo or seal and its name. Here, we use 12-point Universal demi-bold because it looks official and businesslike without being pretentious.

The headings and subheadings, which attract the respondent's eye next, are short phrases that tell what each part of the questionnaire is about. They stand out in 12-point Universal bold and 11-point Universal demi-bold or similar typefaces.

Most of the questionnaire is text containing the instructions, questions, and answer spaces. Here, GAO usually uses 9-point or 10-point Times Roman, Baskerville, Press Roman, or similar type. These are clear, simple, easy-to-read, official-looking typefaces, with good height-to-width ratios, and the 9-point or 10-point size is large enough to read easily yet small enough to keep the questionnaire from getting too bulky.

Once respondents begin to answer the questions, they see the response instructions. These are short texts, usually in parentheses, that tell how to answer—for example, "(Check one.)." Response instructions are usually in an italicized version of the typeface used for the text and are the same size. Like the response

instructions, fill-in-the-blank instructions are in italics and parentheses.

After answering a question, the respondent is frequently directed to another part of the questionnaire by instructions to "skip" or "go to question . . . ." These are usually in 9-point or 10-point bold type. The bold type emphasizes the skip instructions and this helps reduce errors. Occasionally, bold type is used to emphasize a key point in a question or text, such as an important qualifier that might be overlooked. GAO prefers bold rather than underlining because underlining stops eye movement and slows the respondent down.

Next comes the response space—little boxes to check; a row, column, or matrix box to fill in; or sometimes a line for the respondent to write in information. All little boxes for single-response alternatives are left-justified or aligned to the left of the response. The use of square boxes yields fewer errors than circles or other shapes, as does the left justification over right justification, unless a row, column, or matrix format is used. Rows or columns or column-row matrixes are justified to the right, so that they line up with the row and column headings. Boxlines are used instead of leaders because they guide the eye better. All line work should be a half point or 1 point in width. The page looks too dense if the lines are much thicker.

The row headings are in the same type as the text. Sometimes the column headings are in Gothic or similar type. Such typeface can be squeezed more than most others without destroying the letter symmetry of the word and without running the letters together, therefore not interfering with readability. Gothic typeface also reads well for very short passages, but it does not work as well for long passages.

All questions and response alternatives are numbered rather than lettered. These numbers double as codes for information field identifiers for use in data reduction.

Tiny numbers in parentheses to the right of the questions tell the keypunch operator what column to punch in tabulating responses. These column codes are in 5-point or 6-point Gothic italics or similar type. They are not big enough to distract the respondent nor are they too small for the keypunch operator to read.

Shading is used to fill in space that the respondent might confuse with response space. The shading prevents respondents from writing in the space. A row of light shading can also be used to separate rows of text on a long horizontal layout or to guide the respondent across the page.

The form design also makes use of white space. Leaving good margins, top and bottom space, and space between the text columns reduces the clutter, separates key parts of the questionnaire, and makes it look more inviting. Questionnaire designers should try to give the respondent as much white space as possible without expanding the number of pages.

Chapter 14

Preparing the Mail-Out Package and Collecting and Reducing the Data

In addition to developing the questionnaire itself, GAO evaluators generally complete several other tasks, as summarized below:

- develop a computerized mailing list, a cover letter, and other mail-out materials and assemble the mail-out package;
- monitor and edit the returns and conduct follow-ups;
- key in the responses, verify the computer file, and develop the data base.

Preparation of the Mail-Out Package

Before the questionnaire is mailed to potential respondents, a computerized address file has to be developed, a cover letter has to be prepared, and other materials (such as return envelopes) have to be assembled for the mail-out package.

Address Files

Concurrent with designing and testing the questionnaire, evaluators should select the population sample cases for the survey. (See chapter 3.)

It is usually a good idea to send or distribute the packages directly to an individual rather than to rely on intermediaries. Transmittals that rely on intermediaries usually do not work well, and when they go wrong, the survey loses credibility because control of the sample has been lost. In one instance, we gave questionnaires to Veterans Administration hospital administrators to distribute to the staff, and in another we gave them to union leaders to give to their members. Both distributions were incomplete, and both surveys had to be discounted because of poor response rates and uncontrolled sample selection.

It is normal to begin with a hard-copy list of addresses. This list should be reviewed, and careful

attention should be paid to the following matters to ensure that it is current, complete, and accurate:

- spelling and capitalization,
- titles (Dr., Ms., Mr.),
- job titles (as appropriate),
- street addresses with room numbers and apartment numbers (as appropriate), and
- city, state, and zip code.

The revised hard-copy list must now be put into a computerized file. This can be done in several ways. For example, the list can be keyed on tape or disk and entered into the appropriate computer system. The list can also be typed on a word-processing system disk and then transferred to the system, or it can be typed directly into a system file from a remote terminal.

Once the file is in the system, a hard-copy list can be prepared, reviewed, corrected, and case numbered. The address file is in this format:

Mr. John Doe
226 Main St.
Middletown, NY 00000

At this point, a hard-copy log with case numbers should be printed for use in controlling mailed and returned questionnaires.

Transmittal Letter

Because respondents see the cover letter first, their decision to participate in the survey is often made on the basis of the letter's strength. Therefore, the letter should pay attention to the following guidelines, which have been found to increase the likelihood of a reply:

1. Design the mail-out package so the letter is seen first.

2. Have the letter neatly typed to look like a personal, individualized communication rather than printed or xeroxed.

3. Use an official-looking format and style of writing but avoid being impersonal, ambiguous, or unclear.

4. Address the letters to each individual.

5. Explain what GAO is and why it has a legitimate and purposeful role in collecting these data.

6. Without being pretentious, explain that GAO is an important agency working for the Congress.

7. State the purpose of the project.

8. Stress the importance of the project.

9. Relate the project to the respondent.

10. Stress the importance of the answers and the study to the respondent and the nation. If possible, make references to possible benefits to respondents.

11. Tell how and why the respondent was selected.

12. State that the questionnaire can be answered easily and in a short time. Tell truthfully how long it should take to complete the questionnaire.

13. Emphasize the importance of replies from everyone sampled.

14. Ask a favor.

15. When necessary, ensure anonymity or confidentiality and no uses other then those stated.

16. Ask for honest and frank answers.

17. Urge prompt responses.

18. Alert the respondent that there will be a follow-up for those who do not reply.

19. Mention the possibility of a verifying personal interview when appropriate.

20. Provide a name and a phone number in case the respondent needs assistance in completing the form.

21. Express appreciation for the respondent's assistance.

22. Have the letter signed by hand in blue ink by the person with the highest appropriate responsibility. If many letters are to be sent out, have several clerks sign them.

23. Send the package by first-class mail. (The return envelope should also be for first class.)

The pledge of confidentiality is worthy of further discussion. In GAO, we use pledges of confidentiality only when it is essential for meeting the assignment objectives and the data cannot be obtained in another way. We use pledges that individual and organizational names will not be released and that responses will generally be reported in an aggregate form to help increase the response rate and the truthfulness and candor of the respondents. (Before a pledge of confidentially is used, a written justification is prepared and approved by the assistant comptroller general of the division.) For work being done for the Congress, GAO's pledge is approved in writing by

each requester. When GAO's pledges are given, the link between individuals and their responses may be destroyed after all analysis, referencing, and supervisory reviews have been completed. (If a follow-on review is anticipated, it may be necessary to retain the linkage.) The current policy guidance on pledges is included in chapter 6 of GAO's General Policies/Procedures Manual.

Once the transmittal letter has been written, edited, reviewed, and revised, it is ready to be typed into the computer system as a separate file. GAO can run a computer program to produce the transmittal letters by merging each address in the address file with the transmittal letter file. At this point, the letters are ready for signature.

Other Mail-Out Materials

The following materials should be prepared and printed (by printing services) to complete the mail-out package.

1. Preaddressed, postage-paid return envelopes are used to return the questionnaires and are usually addressed to an individual on the project team.

2. Preaddressed, postage-paid postcards for respondents to indicate that they have returned the questionnaire separately are used when the respondents are to remain anonymous to GAO. They tell GAO that the respondents have sent in their questionnaire so we do not follow up on them.

3. Business letter envelopes can be used if the questionnaire is six pages or less. Window envelopes are sometimes used to avoid labels. Large questionnaires and booklets require large envelopes and mailing labels.

4. Occasionally, letters of endorsement from influential people are included in the mail-out package if it is believed they will increase response rates or result in more complete and honest answers. For example, a survey of Navy contractors might be enhanced by including a letter of endorsement from the admiral in charge of contracts or another senior Navy official.

Once all the materials have been gathered together, an assembly line is formed to fold, stuff, seal, and control the mail-out package, using the address list as a control log. These activities are normally done in-house; however, they can also be done by an outside firm when a long lead time is available, the sample is large, and the benefits outweigh the costs.

Data Collection

Essential to a good data collection phase is the monitoring of responses (and nonresponses) and a continuing effort to get the responses. Generally, GAO attempts to attain a response rate of 75 to 95 percent, which is the generally accepted standard of the survey research community.

Monitoring Returns

The address list developed for the mail-out package is an excellent tool for monitoring returns and ensuring that an outcome—a return or a reason for no return—is recorded for each sample unit. This same list will serve as the basis for mailing follow-up materials to nonrespondents. Maintaining this log is very important because it also serves as a control to document the cases that were entered into the computer.

The earliest returns may be undeliverable packages. For each undeliverable, a note should be made on the control list of why the package could not be delivered. Incorrect addresses should be recorded

and new mailings should be prepared when feasible. Other early returns may come from those who were erroneously included in the sample and therefore should not complete the questionnaire. It is important to separate inappropriately sampled units so that both the sample size and the population size can be adjusted. The return of questionnaires should be noted in the control log (usually with the date of return). When anonymity was assured, the returned post cards serve this purpose.

Follow-Up Procedures

Follow-ups can take several forms and can be conducted with varying frequency. For example, a project might begin with an initial mailing and then be followed by one or two follow-ups, using the normal postal system. Final follow-ups might then be conducted, using telephone contacts, mailgrams, or telegrams. Each technique has its advantages in certain situations.

About 3 weeks after the initial mailing, responses will probably drop off each day. They are likely to trail off to a response total of about 30 percent to 50 percent. At this point, a follow-up is needed. Over the years, GAO has found that a single follow-up will bring in about one third to half of the outstanding questionnaires. Thus, we expect that about 3 weeks after mailing the first follow-up, we will have about 50 to 75 percent of our responses. A second mailed follow-up may be helpful at 8 to 9 weeks.

At about the 11-week point, the response rate should be reevaluated in light of project goals. It may be possible to stop or perhaps to try one last follow-up by telegram or telephone. This decision should be based on such factors as (1) the number of outstanding responses (it is practical to call 75, but not 750, nonrespondents), (2) the availability of staff

to make calls, and (3) the availability of resources (telegrams can be costly).

Follow-up letters are prepared and produced in a manner similar to the preparation of the initial transmittal letter. The names of those who responded are subtracted from the mailing list, and a new file is created with the new letter. In the manner described previously, these two files are then merged, a new set of cover letters is produced, and new mail-out packages are assembled and mailed. See figures 14.1 and 14.2 for examples of initial transmittal and follow-up letters.

Figure 14.1: Initial Questionnaire Transmittal Letter

GAO

United States
General Accounting Office
Washington, D.C. 20548

Program Evaluation and
Methodology Division

March 15, 1993

Mr. John Doe
1776 Main Street
Middletown, NY 98765

Dear Mr. Doe:

The U.S. General Accounting Office--an investigating agency of the Congress--is reviewing the effectiveness of your National Guard or Reserve training. We could not undertake this review without first considering the experiences of the people like you who receive this training.

Since it is impossible to talk to each of you in person, we have selected, at random, a sample of people who, like yourself, represent a cross-section of the forces. We are asking each of you to complete a short questionnaire. While this task could take 15 or 20 minutes to complete, your answers are of vital importance to our review and to others like yourself who need this training.

Since the sample represents a very small portion of service personnel, we must hear from everyone or our results will not be representative.

We need your frank and honest answers and we want to make one point clear. Your answers are anonymous and cannot become part of your service record or any other file. There is no identifying information on the questionnaire and nobody can tell how you or any other person answered. We ask only that you return the enclosed post card to tell us that you have completed and returned the questionnaire. Mail this card separately from the questionnaire to preserve your anonymity. We need to know that you have mailed in your reply so we do not burden you with nonresponse follow-up letters. Remember that while your name is not important to our survey, your experiences and opinions are. We cannot make meaningful recommendations without help from you and others like you.

It is essential that you complete the questionnaire and return it in the enclosed envelope within 10 days of receipt.

If you have any problems, please call Mary Glass at (202) 555-9999.

Thank you for your cooperation.

Sincerely,

Jane Doe
Director

Enclosures

Figure 14.2: Questionnaire Follow-Up Letter

GAO

United States
General Accounting Office
Washington, D.C. 20548

Program Evaluation and
Methodology Division

June 15, 1993

John Roe, MD
1492 Center Street
Highville, Pennsylvania 12345

Dear Dr. Roe:

About four weeks ago, we sent you a questionnaire concerning Medicare reimbursements to physicians who treat end-stage renal disease (ESRD) patients. As of today, we have not received your reply. If you have already returned the questionnaire, please excuse this letter and accept our thanks for helping us.

If you have not yet completed the questionnaire, please do so and return it as soon as possible. We need your returned questionnaire to complete our review. Your opinions regarding ESRD physicians' Medicare reimbursements and the Health Care Financing Administration's proposed regulations are of interest to us.

As mentioned in our previous letter, the information you give in the questionnaire will be kept confidential and we will not release it outside of GAO, unless we are compelled by law or required to do so by the Congress. Your responses will be combined with those of other physicians for our report to the Congress. However should the answers of individual physicians be discussed, they will not include information that could be used to identify individual respondents.

We have enclosed another copy of our questionnaire for your convenience. If you have any questions, please call Joe Green at (202) 555-9999.

Thank you for your cooperation.

Sincerely,

Jane Doe
Acting Manager

Enclosure

Editing

As questionnaires are returned, they must be edited before they can be keypunched and entered into the computer system as a file. The editing is done in accordance with a preestablished protocol designed to identify obvious respondent errors and missing data points, make corrections and missing data estimates systematically and appropriately, and make sure the data entry operators can follow and accurately key the responses. The editing process can take weeks to complete, but a team can begin editing as soon as responses are received. Editing should not have to continue more than a short time after the last questionnaire has been received.

To determine whether the responses are adequate, evaluators should look for the following kinds of items:

1. Is the response complete?

2. Did the respondent follow instructions? Skip appropriate questions? Answer appropriate questions? Check the correct number of responses to each question—one or all that apply? Place responses correctly in the response space provided?

3. Is the response sufficiently clear for data entry?

4. Do the open-ended responses provide useful data?

5. Did the respondent just check any response or make wild guesses without consideration by reviewing the consistency of the response pattern?

6. Did the respondent leave a space blank to indicate no or answer a question that should not have been answered?

7. Did the respondent answer clearly? For example, did he or she write "5K" when the correct notation was 5,000?

8. Are there missing values, misplaced answers, unclear responses, and inappropriate answers?

9. Are the missing values clearly distinguishable from the not-applicable answers, skips, or zero values?

10. Are all numbers right justified or positioned to the right to allow for either leading zeros or blank spaces?

11. Were extreme values checked that look inconsistent?

12. Are there responses that are logically inconsistent?

13. Are there mathematical errors in the responses?

After the editors have reviewed perhaps 50 or 100 questionnaires according to these guidelines, they should prepare a written edit protocol that specifies the procedure for making edit changes. For example, this procedure should specify what to do if the respondent checks two alternatives of a set when he or she should have checked only one, and it should specify what items to look at to check for inconsistencies.

Furthermore, some of the edit checks and corrections may be done by the computer after the data have been keyed and loaded on the computer; these include respondent math errors and coding blanks as missing values, no's, or "not applicable." These computer edit protocols should also be specified in the manual edit protocol before the editing starts, to minimize overlooking any edit procedure. Both the

manual and computer edit protocols are used to develop the data entry protocols.

Once edit protocols have been tested, inadequate or obviously incorrect responses must be assigned as missing values or sometimes corrected according to an established protocol for identifying the logically correct answer or adjusted according to further contact with the respondents (usually by telephone). Once the evaluators are satisfied that the responses meet project standards, the data reduction phase of the survey can begin.

Data Reduction

Before the data can be analyzed, they must be moved from hard-copy form (the questionnaire) into a computerized data file that accurately reflects the hard-copy data. This process begins with keying the data onto a medium and in a format that the computer can read.

Keying

Keying for GAO questionnaires is normally done by an outside contractor. Nearly always, the contractor keys from one of two sources—the questionnaires themselves or a coding sheet usually laid out in an 80-column card image format and prepared by the project team. Many GAO questionnaires are coded on an 80-column card format for ease of editing, not for ease of keying. The keying is generally done onto a tape or disk (not cards) that can readily be entered into the computer system as an unedited raw data file. Every data entry key stroke is first verified by the contractor. Keying instructions unique to the individual job are provided to the keyers for guidance. These should be written in conjunction with the manual and computer edit protocols so none of the edit or keying considerations are left to chance. Two of the evaluators' primary tasks are to ensure that the questionnaires given to the keyers are keyed and that

all original questionnaires are returned—a control function.

Keyed Data Verification

In a first, short, but necessary step, the tape or disk containing the unedited raw data file is loaded in the computer system. Once loaded in the computer system, the unedited raw data file can be converted to hard copy, in order to verify for a second time that the computer file accurately reflects the contents of the questionnaires. For GAO projects, at least 99 percent of the keyed strokes must be correct to be considered accurate. When unacceptable error rates are found, the data are keyed in again.

Rather than verify the entire file, every question in every questionnaire of a sample of questionnaires can be verified. This is a cluster sample. How large should the sample be? It should be large enough to statistically ensure, at the 95-percent confidence level, that the data entry error rate is not more than 1 percent (1 plus or minus 0.4 percent). This often amounts to a 10-percent sample of cases—for a typical job of about 400 questionnaires and a typical questionnaire of about 250 characters. However, if the number of questionnaires or the number of characters per questionnaire is smaller or larger than the typical case, it is necessary to sample more than or less than the 10 percent, respectively. Also, if a greater precision is needed (error rate less than 1 percent) or if no error rate is permitted, then much larger samples or all the data must be verified by the evaluation team. Table 14.1 shows the percentage of questionnaires that might be sampled for a 1-percent error rate. However, the table is only a rule of thumb. A sampling expert should be consulted to determine the appropriate sample for keyed data verification. The verification process works best when two evaluators work together; one reads from the

questionnaire while the other views the printed computer file.

Table 14.1: The Percentage of Questionnaires That Should Be Randomly Sampled to Determine the Keypunch Error Rate

Number of questionnaires per keypunch job	Characters per questionnaire		
	1-99	100-300	More than 300
1-29	100.0	100.0	100.0
30-99	50.0	40.0	25.0
100-499	10.0	10.0	7.5
500-999	7.5	5.0	5.0
More than 1,000	5.0[a]	5.0[a]	5.0[a]

[a]The maximum number of questionnaires in the sample should be 384.

Even when an acceptable error rate is found, errors noted during the review should be corrected for the sampled cases. In addition, noted error patterns should be investigated. For example, assume the reviewers note (frequently a judgment call) that the keyer misinterpreted the responses to a question. Then all the responses to that question should be verified and corrections made. An additional edit should be made on all questions that can take on only a limited number of values. For example, a yes-no question may have values limited to 1 or 2, and a question asking about an item's cost may be known to have an upper limit of $10,000. A computer program that checks for out-of-range values should be run and corrections made.

After this process has been completed, an edited raw data file is available that can be used in the initial steps of the analysis phase, as discussed in the next chapter. It is also important to note that the data

verification noted above is a minimum protocol and that each GAO division may have additional, more specific and rigorous requirements.

Chapter 15

Analyzing Questionnaire Results

Analysis Plan

As noted earlier, a data analysis plan should have been developed as part of the evaluation design, after the questionnaire and sample have been developed but before any data are collected. Thinking through the data analysis may cause evaluators to reconsider their data-collection plan or even the evaluation questions themselves. In planning the data analysis, they might realize, for example, that they need additional data that they had not thought of before.

An analysis plan also forces evaluators to decide what kind of findings they do and do not need to complete the evaluation. This process is important, because it is very easy to overburden the study with unnecessary analyses. Since most standard analysis packages can provide millions of analyses that would take many years to interpret, evaluators have to run the analysis; otherwise, it will run them. Also, unplanned analysis can result in fishing or data dredging—that is, the running of analyses without regard to a design or preconceived reason, just to see what will turn up. However, while a plan helps, evaluators cannot always predict relations that might emerge in exploratory analysis.

The selection of analysis techniques and the variables to be analyzed will be determined to a large extent by the evaluation questions and the design requirements. Evaluators also need to make sure that their statistical analysis software routines can satisfy these requirements. For example, can they handle the size, number, specification, and measurement of the variables? And can they do the analyses required? Furthermore, the choice to do certain kinds of analysis often requires a respecification of the variables, measures, and variable relationships.

Later, when the analysis begins, the evaluators will know how adequate their planning and data collection have been. If the measures were properly

defined, relevant, and sound, and if the data relationships turn out as hypothesized, then the analysis will proceed as planned. However, projects are rarely perfect—there usually are some gaps in the planning and problems in the data collection. Measures are not always properly specified. Some important data may not be collected and some of the data that are collected may be irrelevant or unsound. Evaluators need then to modify the analysis plan, perhaps by scaling back the effort, expanding it to cope with unexpected developments, selecting methods to handle missing data, or exploring different ways of answering the evaluation questions. Regardless of departures from the original plan, however, the analysis must still proceed logically and step by step from very simple analyses to a limited number of more complex analyses.[1]

Item Responses and Univariate Analysis

The first step is to go just a short way beyond the raw data on questionnaires by producing column, row, group, and subgroup tabulations and percentages, often called a "code book." The code book tells how people answered each item on the questionnaire by frequencies and percentages for each possible response category. Going one step further in the data analysis, evaluators can compute descriptive statistics and other indicators that help describe the frequency distributions.

Bivariate Analysis and Comparison of Two Groups

Comparisons between groups of respondents can be made. If evaluators want to study the relationship between two variables, they use correlational techniques, which show that a change in one variable is associated with a change in another. For example,

[1]The point of this chapter is to provide guidelines for developing a data analysis strategy. For a more detailed discussion of the quantitative techniques to implement this strategy, see U.S. General Accounting Office, Quantitative Data Analysis: An Introduction, GAO/PEMD-10.1.11 (Washington, D.C.: June 1992).

we might want to determine whether the performance of the Federal Aviation Administration's flight-station service specialists decreases appreciably with age. We would plot the performance scores of specialists of various ages and see whether performance is related to age. We might use an analytic technique such as correlational analysis, which shows the degree to which two variables are related. Or we might compare the differences between two groups rather than the association between variables. For example, we might compare the performance of younger specialists with that of older specialists. Other primary analysis techniques would include cross tabulations, chi-square comparisons, "t" tests, and analyses of variance.

Multivariate Analysis and Comparison of Multiple Groups

This level of analysis is used when what is wanted is a look at the associations between more than two variables or at differences between more than two groups. For example, we might want to study the effect of age and experience on Federal Aviation Administration specialists' performance or the effect of age, experience, training and education, and recency of training and education all together. Here, we could use such multivariate techniques as partial correlations, multiple regression analysis, and factor analysis. We could also compare performance by looking at the differences between groups that have varying levels of each trait (older and experienced, younger and experienced, older with limited experience, younger with limited experience, and so on). We might use such techniques as multiple analysis of variances, discriminant analysis, linear structural relations, or log-linear analysis.

Choice of Analysis Methods

The choice of data analysis methods depends largely on the evaluation questions and subject matter under study and on the type of variables and what levels of

measurement they satisfy. For example, if we had a question about whether the performance of Federal Aviation Administration specialists is different at different ages, and if we had reason to believe that performance was related to age and little else, a simple correlational analysis would reveal the degree of the relationships. But the matters GAO studies are usually more complicated than this, so we would expect other variables such as experience, education, training, and recency of education and training to be related to performance. We would need then to perform multivariate analysis in order to determine the relationships of the variables. Likewise, it might be important to compare performance across several groups rather than to confine the analysis to simple contrasts between pairs. The more complex analyses should usually be undertaken only after the results of simpler analysis have been examined.

Sometimes evaluators have a choice between using associations and using group differences, and sometimes they do not. The shape of the data distribution, the measurement scales, and the plots of the functional relationship between the variables may rule out the use of correlation techniques. For example, sometimes we have to study group differences because the distribution of the observations is not normal; we could not then use certain correlational statistics. Correlational techniques are also inappropriate when the variables are scaled with ordinal data and when the relationships under study are not linear—that is, the plot between the variables cannot be transformed into a straight line. It is important to realize that correlational techniques cannot by themselves be used to show causality.

Because questions about cause and effect are sometimes posed, we must note that special designs such as nonequivalent comparison groups, regression

discontinuity, and interrupted time-series are usually necessary for establishing causality. The logic of the evaluation design, not the analytic technique, is crucial in drawing inferences about causality.

Chapter 16

Adaptations for the Design and Use of Telephone Surveys

Telephone surveys are occasionally used in GAO assignments as another method for collecting structured data. However, there are differences between mail and telephone surveys, and they cannot be used interchangeably. To help evaluators appropriately use telephone surveys, we discuss the principal advantages and disadvantages of this methodology and the design requirements, adaptations, and administrative considerations in this concluding chapter.

Advantages and Disadvantages of Telephone Surveys

We use telephone interviews at GAO when time is essential. With sufficient staffing, a telephone survey can be completed in days as opposed to weeks for a face-to-face interview and months for comparable mail surveys. For some assignments, they are the only feasible approach. For example, in one audit we were required to estimate all the homeless children in shelters nationwide on a given day. With prior arrangements and scores of callers, we called a national sample of shelters to get the count. Given the nature of the shelter environment and the prohibitive costs of face-to-face interviews, no other method would have been possible. While not as cheap as mail surveys, telephone interviews cost much less than face-to-face interviews. Telephone surveys may cost between $40 and $75 per case as opposed to hundreds of dollars for personal interviews and a few dollars per case for mail questionnaires.

Telephone surveys have certain advantages and disadvantages. For one thing, people seem to prefer mail surveys and personal interviews. For another thing, while telephone surveys are less sensitive to certain design problems, they are more sensitive to others. For instance, telephone instruments may be less sensitive to design flaws likely to cause primacy bias errors but more likely to be affected by recency bias errors. In addition, telephone responses are less

complete and less accurate than mail or face-to-face responses. Many people do not like to talk on the telephone for long periods of time, because they do not like tying it up. They feel pressured to answer. They are more likely to answer in extremes. They answer from the top of their head and truncate their memory search earlier than they do in other modes of data collection. They are also more likely to acquiesce, guess, or give any answer or an easy answer or the same answer to all than in some of the other methods of interviewing.

However, telephone interviews are an important and valid means of collecting data. In fact, the private sector, partly because it does not enjoy GAO's high mail response rates, relies very heavily on telephone surveys. Many government agencies and private sector businesses that must deal with the public sector also depend on the telephone because they sometimes have difficulty obtaining current and accurate address lists. Hence, these comments are not meant to discourage the use of telephone surveys. Telephone surveys are an important data collection method available for the evaluators' use. But what this means is that telephone surveys have to be very carefully crafted and adapted to the telephone medium.

Design Guidelines

1. Minimize instrumentation errors. As a medium, the telephone magnifies the effect of certain design problems and minimizes the effect of others. In this section, we consider design problems that have a much greater effect on telephone surveys than on the mail questionnaire or on face-to-face interviews. This sensitivity is partly caused by the telephone medium's not having the added cues inherent in the mail and face-to-face interviews. In the mail survey, all the information is presented simultaneously. The respondent can easily skip back and forth and use the

context for help. In face-to-face interviews, both the interviewees and the interviewers use the paralanguage (gestures, facial expressions, and so on) to help understand. Thus, the interviewee can ask for help without actually asking and the interviewers can often tell if their messages are understood. Also, in face-to-face interviews, this advantage can be complemented by the use of visual cues such as show cards.

The telephone medium worsens the effects of design problems in clarity, construct development, language level selection, qualifications, question format selection, response categorization, question bias, facilitating memory recall, and minimizing undesired ordering and recency effects. However, on the positive side, it may be that design flaws in accounting for primacy and undesirable context effects will have less effect when the question is asked on the telephone. For guidelines to resolve these types of flaws, the reader is referred to chapters 2, 4, and 6-12 of this publication and to Using Structured Interviewing Techniques.[1] These materials provide sufficient guidelines to correct most design flaws. However, in the rest of this section, we add some design notes on clarity, facilitating short-term mental processing and long-term retrieval, guiding the line of questioning, minimizing the cognitive tasks, minimizing recency effects, setting up reasonable interview time lengths, and pretesting.

2. Stress clarity. Follow the guidelines on clarity specified in chapter 6 to the letter. Pay particular attention to the following suggestions. Use conversational English. Write with short and simple syntax. Limit the sentences or syntactical structure to 20 or 25 words. When possible, use familiar words. Use concrete words if abstract words can be avoided.

[1]U.S. General Accounting Office, Using Structured Interviewing Techniques, GAO/PEMD-10.1.5 (Washington, D.C.: July 1991).

Use words that are easy to picture or imagine. Make sure all the words you use and the way you use them have a single meaning or a very limited number of meanings. Make sure all important qualifications are stated in such a way that they will be noted and understood.

3. Write to facilitate both short-term processing and long-term memory retrieval. Follow the guidelines specified in chapter 8 and carefully consider the following. Limit to 25 words or less each idea or unit of information that is to be kept in the listener's head long enough for the higher cognitive process to work. This is because respondents have trouble comprehending speech that is spoken faster than 100 words a minute or 25 words per 15 seconds. The immediate memory span or processing capability, the amount of time most people can keep information in their heads without losing it or subjecting it to additional mental processing, is 15 seconds. Limit the higher-level cognitive tests to steps or responses that require no more than 15 or 30 seconds to answer. If you cannot do this, alter the question or the script so that there is interviewer or interviewee feedback or interaction at least every 45 seconds. Respondents feel pressured to answer quickly to alleviate the silence between question and answer during a telephone interview. Forty or 50 seconds of silence is often too much pressure. To resolve it, many will cut the quality of their mental processing in order to answer more quickly.

4. Guide the line of questioning. Respondents provide more complete and accurate answers if they can anticipate the line of questioning and the information they must retrieve. Telling the respondents where you are going or providing transitions so that the next questions can easily be inferred or anticipated helps them to some anticipatory cognitive processing that improves their ability to answer. In some of the

alternative methodologies, the respondents can use the context of the instrument to warn them of what is coming next. But telephone interviews are devoid of such context cues.

5. Decompose the cognitive tasks. Make sure all complex and difficult comprehension tasks are broken down into small steps and that these steps form discrete, complementary, and logical operations. Unlike with other media, respondents do not have a visual format of the problem. They must keep all the rules, conditions, and qualifications in their heads. Also, remember that many feel they must answer quickly. If the task appears difficult, the interviewee often resorts to inefficient and error-prone heuristics and strategies. This does not mean that we cannot use the telephone to audit complex issues. It just means that we must break down a complex inquiry into smaller, logically ordered operations.

6. Minimize recency effects. Perhaps the biggest difference between telephone surveys and alternative methods is a pronounced recency bias. That is, alternatives and conditions presented in the latter part of the question will be remembered best and, hence, are more likely to be chosen. One way to mitigate this effect is to limit the alternatives to seven choices or, if the choices are more complex, to five or fewer.

7. Minimize the tendency to extremes. Telephone respondents appear to be more likely to answer in extremes than their counterpart mail and face-to-face interview respondents. To minimize this tendency, use the techniques discussed in the latter part of chapter 4 on intensity scale formats. That is, use a branching format with a middle alternative whenever possible. Also, be careful to use well-anchored, equal-appearing intervals in the response scales.

8. Keep the interview short. For a variety of reasons, some people get uncomfortable during extended telephone interviews. If the interview goes over half an hour, they feel somewhat stressed. Repetitions become tedious. They are more likely to acquiesce, to guess, to cut their memory or cognitive tasks short, to answer from the top of their heads, to select extremes, or to use other forms of shortening their responses.

9. Pretest the interview under realistic conditions. Follow the procedure described in chapter 12 with the following exceptions. Administer the pretest over the telephone so the interviewer is not in the presence of the respondent. Have an observer (a person different from the interviewer) be present to observe the respondent and take notes and response times as he or she would in a normal pretest procedure. The observer should be able to hear the interviewer over an extension. If this is not possible, the observer should at least have a copy of and be familiar with the script in order to follow the interview. The pretest debriefing should be conducted as if it were a normal pretest.

Here are some cautions concerning the dual administration of telephone and mail surveys. On occasion, evaluators may consider using mail, telephone, and face-to-face interview methods to administer the same instruments. For example, they might want to use a telephone survey to complete the last follow-up of a mail survey because the telephone methods require less calendar time. Another example is the use of a mail survey to contact part of a telephone survey population that could not be reached because of unlisted numbers, duty overseas, or other reasons. While this can sometimes be done without compromising the survey, it is usually not a good idea. As we can infer from the preceding chapters on mail surveys and the previous discussion

on telephone surveys, these methods have different effects and can produce different results.

Consider these differences. Mail surveys are more prone to primacy bias and contextual cues than telephone surveys. Telephone surveys show more recency bias, social desirability bias, and tendencies toward extremes than mail surveys. Furthermore, these differences become even greater as the cognitive requirements of the subject matter become more difficult.

This does not mean that we can never use a mixed mode. We have in fact used it successfully in several studies. However, in each of these cases, we were careful to plan and design for a mixed mode administration. In addition, if mixed approaches are to be used, it is imperative that the survey responses be tested to rule out or account for mode differences.

Administration

Telephone survey administration requires an advance letter, a contact log, a trained staff, a monitoring procedure, and, if possible, computer assistance.

Advance Letter

The purpose of the letter is to alert respondents as to who you are, why you are calling, and when you expect to call. This establishes your legitimacy prior to the call, thus breaking down some of the respondents' reluctance. It also facilitates the interview because the respondents can refresh their memory, consult records, and sometimes have the necessary information at hand. The letter minimizes the chances of contacting the wrong person and helps increase first-call contact rates because the respondent is aware of your interview schedule. While advance letters are not essential, and you can obviously conduct an interview without one, they

have a very great effect on facilitating the data collection.

Telephone Log

All interviewers should keep a telephone log; it becomes part of the data collection record along with the completed interview. The major purposes of the log are to keep a nonrespondent record, to provide data to make sampling adjustments, to facilitate call backs, and to identify reluctant respondents. In GAO's experience, an interviewer can complete from 6 to 12 half-hour interviews a day. We use the log because it is rare that an interview is completed on the first call. It usually takes at least three calls to get a completed interview. The time and days of the calls are varied so as to increase the chances of getting a contact. If a contact for other than the respondent is made, then the caller should verify the respondent's identifying information and seek referral information. For example, ask for the best time, day, date, and number to reach the respondent or other numbers or other people who may help locate the respondent. Referrals may even help in the search for proxies, if this option was part of the design. Finally, the log should state the status of the interview. That is, was the interview completed, partially completed, or refused? Are more calls planned? (Give a justification, if not.) Are there call-back appointments? (Give time and place, if yes.)

Telephone Interview Log Entries

1. Case number
2. Sample strata number
3. Name
4. Title
5. Address
6. Work number

7. Home number

8. Other numbers

9. Call try, 1st, 2nd, 3rd, etc.

10. Date and time of call

11. Contact, yes or no (if "no," why no: busy, no answer, disconnected, wrong number, no other number, moved, or other)

12. Contact party (if yes):

- respondent, yes or no
- other, yes or no

If other, name, title, and number

13. Referral information:

- name and number of referral
- names, titles, and numbers of possible locators and date and times for best chances of contact

14. Interview status:

- eligible, yes or no
- complete, yes or no
- partially complete, yes or no
- refusal, yes or no
- call back, yes or no
- if no, justification
- call-back appointment, yes or no
- date and time of appointment, if yes
- other status (specify)

15. Proxy information, if relevant:

- name, title, address, and number of proxy

- justification for proxy use

Refusals, No Contacts, and Proxies

As with all surveys, decisions have to be made with refusals, no contacts, and proxies. In telephone surveys, if refusals are numerous enough to be of concern, they are often referred to a more experienced or different interviewer. After a week or two, this interviewer again calls the persons who refused and attempts to persuade them to complete the interviews. Interviewers experienced with "conversions," as they are called, can usually convert from one third to half of the refusals.

If the refusals fail to convert, then the interviewer tries to see if an alternative method would be acceptable. If not, the interviewer tries to get a limited response or reasons for refusing. Alternative methods are mail questionnaires and face-to-face interviews. However, if alternatives are used, they should be analyzed for media and reluctant respondent effects before they are included in the data base. A limited response might be the answer to one or two questions if there are a few questions that are much more important than the rest. These questions should be reviewed before they are included in the data base because they were taken out of interview context. Such interviewees should be considered reluctant respondents. Also, the interviewer should attempt to find out the respondent's reasons for refusing. In addition to statements like "too busy," "not interested in the problem," "don't give interviews," "don't know who you are," valid reasons why the respondent should not have been part of the population are sometimes given. If the reasons are valid, then such cases can be dropped from the sampling group. The refusal group should also be analyzed, if possible, for characteristic differences from the respondent sample.

A case is labeled "no contact" if the respondent could not be contacted after several tries (usually seven) and little or no referral information is available. As with the refusal group, the no-contact group should be analyzed to see if its members have left the population and if they are different from the respondents. Very often the no-contacts should not have been considered part of the sampling frame. Possible reasons for no-contacts are that they left the area, changed jobs, retired, died, discontinued or changed or reallocated operations or responsibilities, gone out of business, and so on. The analysis of the no-contact group characteristics is also important. The extent to which they are different from those of the respondents may affect the external validity.

Sometimes it may be possible or even more appropriate to substitute a proxy for a selected respondent. For example, the interviewer may have found out that the respondent's responsibilities were transferred to another department or that his or her responsibilities were shared by another co-worker or supervisor. However, this substitution should be justified before it is implemented.

Training, Monitoring, and Computer Assistance

All interviewers should be trained and rehearsed in the administration of the interview. Most important, they should be trained to speak at no more than 100 words per minute. If the interview is not standardized, then the interviewer should follow the probes denoted in the script as well as other problem areas that may have been signaled by the respondent's paralanguage. Examples are changes in pitch, enunciation, speech rate, and word trailing. Methods for this type of interview are included in chapter 12. Methods for training in standardized interviews are described in Using Structured Interviewing Techniques. These references stress the importance of interview standardization and interviewer training.

Each contact should be interviewed according to the prepared script and asked the same questions in the same way.

The telephone interviewing should be monitored. Most commercial telephone interviewing operations have a centralized system in which monitors can hear both parties as the interviews are being conducted. Other systems of monitoring allow for more limited monitoring such as on-line sampling of two-way conversations, supervisor monitoring of just the interviewer's conversation, or a recording of the two-way conversation. While the centralized system is superior to the other alternatives, monitoring is very important. It is very important that all interviewers maintain the same level of enthusiasm and professionalism necessary to keep the interview going and on track and avoid feedback that will have untoward or biasing effects.

Often telephone interviews are programmed into a computer-assistance package that facilitates response recording and the administration of the interview. These methods are described in Using Structured Interviewing Techniques.

Bibliography

Baddeley, A. D. The Psychology of Memory. New York: Basic Books, 1976.

Belson, W. The Design and Understanding of Survey Questions. London: Gower, 1981.

Biderman, A. D. (ed.). An Inventory of Surveys of the Public on Crime, Justice and Related Topics. Washington, D.C.: U.S. Government Printing Office, 1972.

Biemer, Paul, et al. (eds.). Measurement Errors in Surveys. New York: John Wiley and Sons, 1991.

Bradburn, M. N., and S. Sudman. Response Effects in Surveys. Chicago: Aldine, 1974.

Converse, Jean, and Stanley Presser. Survey Questions: Handcrafting the Standardized Questionnaire. Beverly Hills, Calif.: Sage Publications, 1986.

Deming, W. W. Sampling Design in Business Research. New York: John Wiley and Sons, 1960.

Dillman, D. A. Mail and Telephone Surveys. New York: John Wiley and Sons, 1978.

Erdos, P. L. Professional Mail Surveys. New York: McGraw-Hill, 1970.

Flesch, R. Say What You Mean. New York: John Wiley and Sons, 1974.

Frey, S. H. Survey Research by Telephone. Beverly Hills, Calif.: Sage Publications, 1983

Groves, R. M. Survey Errors and Survey Costs. New York: John Wiley and Sons, 1989.

Groves, R. M., et al. Telephone Survey Methodology. New York, John Wiley and Sons, 1988.

Krosnock, J., and L. Fabrigar. "Cognitive Perspectives on Survey Questionnaire Design." Manuscript, Ohio State University, Columbus, Ohio, 1991.

Lockhart, D. C. (ed.). Making Effective Use of Mailed Questionnaires. San Francisco: Jossey-Bass, 1984.

Moser, C. A., and Graham Kalton. Survey Methods in Social Investigation, 2nd ed. London: Heineman, 1971.

Oppenheimer, A. N. Questionnaire Design and Attitude Measurement. New York: Basic Books, 1966.

Payne, S. L. The Art of Asking Questions. Princeton: Princeton University Press, 1951.

Rosenberg, M. The Logic of Survey Analysis. New York: Basic Books, 1968.

Rossi, Peter, James Wright, and Andy Anderson (eds.). Handbook of Survey Research. New York: Academic Press, 1983.

Schuman, H., and S. Presser. Question and Answers in Attitude Surveys. New York: Harcourt Brace Jovanovich, 1981.

Sudman S. Applied Sampling. New York: Academic Press, 1976.

Sudman, S., and M. N. Bradburn. Asking Questions. San Francisco: Jossey-Bass, 1982.

Sudman, S., and M. N. Bradburn. Response Effects in Surveys. Chicago: Aldine, 1974.

Turner, C. F., and E. Martin (eds.). Surveying Subjective Phenomena, Vols. 1 and 2. New York: Russell Sage Foundation, 1984.

U.S. General Accounting Office, Pell Grant Validation Imposes Some Costs and Does Not Greatly Reduce Award Errors: New Strategies Are Needed, GAO/PEMD-85-10. Washington, D.C.: September 1985.

Warwick, D. P., and A. C. Lininger. The Sample Survey: Theory and Practice. New York: McGraw-Hill, 1975.

Glossary

Anchors	Anchors are items that serve as reference points from which other items in the series or other points in the scale are judged or compared. For example, the opposite ends or poles of a scale identify the extremes so all values within the scale are either greater or less than one of these extremes. Also, the scale midpoint serves as an anchor in that it either divides the scale into categories or quantifies the half value.
Attribute	A characteristic that describes a person, thing, or event. For example, being female and male are attributes of persons.
Bias	Words, sentence structure, attitudes, and mannerisms that unfairly influence a respondent's answer to a question. Bias in questionnaire data can stem from a variety of other factors, including choice of words, sentence structure, and the sequence of questions. Both interviewer and instrument bias can exist.
Bivariate Analysis	An analysis of the relationship between two variables.
Confidence Level	The level of certainty to which an estimate can be trusted. The degree of certainty is expressed as the chance that a true value will be included within a specified range, called a confidence interval.
Construct	A concept that describes and includes a number of characteristics or attributes. The concepts are often unobservable ideas or abstractions, such as "community," "well-being," performance," or "democracy," that are represented by observable measures.

Estimation Error	The amount by which an estimate differs from a true value. This error includes the error from all sources (for example, sampling error and measurement error).
Judgment Sample	A sample selected by using discretionary criteria rather than criteria based on the laws of probability.
Measure	A neutral concept that determines which data will be collected. The chief methodological concern in developing a useful measure is its validity.
Measurement	A procedure for assigning a number to an object or an event.
Measurement Error	The difference between a measured value and a true value.
Multivariate Analysis	An analysis of the relationships between more than two variables.
Nonrespondent	A person who fails to answer either a questionnaire or a question.
Operationalization	A process of describing constructs or variables in concrete terms so that measurements can be made.
Precision	The exactness of a question's wording or the amount of random error in an estimate.
Reliability Assessment	An effort required to demonstrate the repeatability of a measurement—that is, how likely a question may be

	to get consistently similar results. It is different from verification (checking accuracy) or validity (see Validity Assessment).
Response Style	The tendency of a respondent to answer in a specific way regardless of how a question is asked.
Sampling Error	The maximum expected difference between a probability sample value and the true value.
Scale	A set of values with a specified minimum and maximum.
Standardized Question	A question that is designed to be asked or read and interpreted in the same way regardless of the number and variety of interviewers and respondents.
Unit of Analysis	The class of elemental units that constitute the population and the units selected for measurement; also, the class of elemental units to which the measurements are generalized.
Univariate Analysis	An analysis of a single variable.
Validity Assessment	The procedures necessary to demonstrate that a question or questions are measuring the concepts that they were designed to measure.
Variable	A logical collection of attributes. For example, each possible age of a person is an attribute, and the collection of all such attributes is the variable age.

Verification	An effort to test the accuracy of the questionnaire response data. The concern is uniquely with data accuracy and deals with neither the realiability nor the validity of measures.

Papers in This Series

This is a flexible series continually being added to and updated. The interested reader should inquire about the possibility of additional papers in the series.

The Evaluation Synthesis. GAO/PEMD-10.1.2.

Content Analysis: A Methodology for Structuring and Analyzing Written Material. GAO/PEMD-10.1.3, formerly methodology transfer paper 3.

Designing Evaluations. GAO/PEMD-10.1.4.

Using Structured Interviewing Techniques. GAO/PEMD-10.1.5.

Using Statistical Sampling. GAO/PEMD-10.1.6.

Developing and Using Questionnaires. GAO/PEMD-10.1.7.

Case Study Evaluations. GAO/PEMD-10.1.9.

Prospective Evaluation Methods: The Prospective Evaluation Synthesis. GAO/PEMD-10.1.10.

Quantitative Data Analysis: An Introduction. GAO/PEMD-10.1.11.

Ordering Information

The first copy of each GAO report and testimony is free. Additional copies are $2 each. Orders should be sent to the following address, accompanied by a check or money order made out to the Superintendent of Documents, when necessary. Orders for 100 or more copies to be mailed to a single address are discounted 25 percent.

Orders by mail:

U.S. General Accounting Office
P.O. Box 6015
Gaithersburg, MD 20884-6015

or visit:

Room 1000
700 4th St. NW (corner of 4th & G Sts. NW)
U.S. General Accounting Office
Washington, DC

Orders may also be placed by calling (202) 512-6000 or by using fax number (301) 258-4066.

United States
General Accounting Office
Washington, D.C. 20548

Official Business
Penalty for Private Use $300

First-Class Mail
Postage & Fees Paid
GAO
Permit No. G100