

GAO

United States General Accounting Office

Program Evaluation and
Methodology Division



July 1986

Developing and Using Questionnaires

Transfer Paper 7

036220 / 130587

PREFACE

In the last two years, about one of every five reports on GAO evaluations and audits used mail questionnaires to collect data. Although questionnaires can be cost effective when they are appropriate to the overall study questions posed, they are not always the method of choice for answering all types of study questions.¹ But even when they are the method of choice, questionnaires, like other data collection methods, are vulnerable to error. To develop and apply one well requires a certain amount of expertise. Given, then, that the use of questionnaires is widespread within GAO and that their yield can be maximized by capitalizing on past experience, we thought it might be helpful to prepare a transfer paper on this topic for GAO staff.

The present document summarizes the most important principles and procedures used in developing, writing, and analyzing effective questionnaires. Its purpose is not only to explain this process so that GAO evaluators can take a more active role in questionnaire development but also to demonstrate some of the specialized skills and kinds of professional help that may be needed to construct and use a questionnaire optimally. We do not expect that GAO evaluators, after reading this paper, will become experts in the preparation of questionnaires. Instead, we want to provide enough information about questionnaires to enable evaluators to (1) understand the activities involved, (2) work effectively with measurement specialists in the development of questionnaires, (3) in some cases, assist in all or part of the questionnaire development tasks, (4) become aware of the principles of questionnaire development, and (5) judge the quality of the data collection effort as a whole.

This document does not by itself address every aspect of questionnaire development and use. Therefore, we recommend that, before GAO project staff decide to employ a questionnaire on a particular job, they request help from either the design and methodology technical assistance group in the GAO division that is programming the assignment or the measurement assistance staff in GAO's Program Evaluation and Methodology Division (PEMD).

Developing and Using Questionnaires is organized according to the sequence of tasks necessary to produce a data collection instrument and to collect the data. Chapter 1 presents an overview of the process and chapter 2 discusses advantages and disadvantages of questionnaires. The two following chapters--

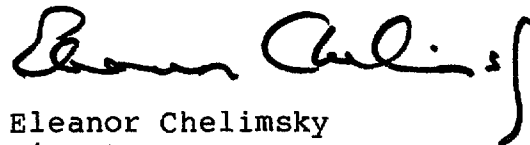
¹The questionnaire can be used as a method for collecting data with several evaluation strategies. For a discussion of these strategies, see Designing Evaluations, PEMD methodology transfer paper 4 (Washington, D.C.: U.S. General Accounting Office, July 1984).

that is, chapters 3 and 4--focus on the initial phases of questionnaire development--planning the questionnaire and sample design. Then eight chapters (chapters 5-12) address questionnaire writing (including question formatting); avoiding inappropriate questions by means of pretesting; clarity of language; answer choices; question bias; response bias, measurement error, and scales; and organizing a logical line of inquiry. The concluding chapters (chapters 13-16) explain quality assurance procedures, form design, mail-out packages and data collection and reduction, and analysis of the questionnaire results. In short, we have organized this transfer paper so as to present the work that needs to be done and at the same time explain the difficulties and problems evaluators are likely to encounter.

The material we present here is not new but it is both comprehensive and specifically oriented to GAO studies. It is based on the work of leading practitioners in the field, on a review of the literature, and on our own trial-and-error experiences with the questionnaires GAO has used in over 1,000 evaluations. Policy guidance for use of questionnaires may be found in the GAO Project Manual and the GAO General Policy Manual.

PEMD has issued and is developing additional transfer papers to keep GAO project staff abreast of concepts and techniques that are useful to both auditors and evaluators. The reader is referred to Designing Evaluations (and its workbook), Causal Analysis, Content Analysis, Using Structured Interviewing Techniques, and Using Statistical Sampling for further information on those subjects.

The authors of this paper are Brian Keenan, Principal Survey Methodologist of PEMD, and Marilyn Mauch, formerly of PEMD, now with the National Security and International Affairs Division. Readers of this paper are encouraged to convey their questions or comments to Brian Keenan (202-275-7329) or to me (202-275-1854).



Eleanor Chelimsky
Director

C o n t e n t s

| | <u>Page</u> |
|--|-------------|
| PREFACE | 1 |
| CHAPTER | |
| 1 WHAT WE NEED TO KNOW ABOUT QUESTIONNAIRES: AN OVERVIEW | 7 |
| The role of questionnaires in GAO evaluations | 7 |
| What we have to know to use questionnaires | 8 |
| 2 ADVANTAGES AND DISADVANTAGES OF MAIL QUESTIONNAIRES | 18 |
| Choosing the data collection method | 18 |
| Comparative advantages of mail questionnaires | 20 |
| Comparative disadvantages of mail questionnaires | 23 |
| 3 PLANNING THE QUESTIONNAIRE AND DEVELOPING THE MEASURES | 29 |
| Reviewing the evaluation design and specifying the variables | 29 |
| Operationalizing the variables | 30 |
| Identifying standards | 33 |
| Relating the measures to the unit of analysis | 34 |
| Initial analysis and validation plans | 35 |
| 4 DEVELOPING THE SAMPLE FOR DATA COLLECTION | 37 |
| Statistical sampling | 37 |
| Survey universe | 37 |
| Confidence and precision | 41 |
| Sampling error | 42 |
| Calculating the sample size | 43 |
| Nonstatistical sampling | 44 |
| 5 FORMATTING THE QUESTIONS | 47 |
| Open-ended questions | 47 |
| Fill-in-the-blank questions | 48 |
| Yes/no questions | 50 |
| Implied-no choices | 52 |
| Single-item choices | 53 |
| Enfolded formats | 54 |
| Free choices | 55 |
| Multiple-choice format | 56 |
| Ranking questions | 56 |
| Rating questions | 58 |
| Guttman format | 61 |

CHAPTER

Page

| | | |
|---|--|----|
| | Likert and other intensity scale formats | 62 |
| | Semantic differential format | 66 |
| | Paired comparisons and constant referent comparisons | 66 |
| 6 | AVOIDING INAPPROPRIATE QUESTIONS: THE IMPORTANCE OF PRETESTING | 68 |
| | Questions that cannot or will not be answered accurately | 68 |
| | Questions that are not geared to respondents' depth and range of information and perceptions | 70 |
| | Questions that are not relevant to the evaluation goals | 72 |
| | Questions the respondents perceive as illogical or unnecessary | 73 |
| | Questions that require unreasonable effort to answer | 73 |
| | Embarrassing questions | 74 |
| | Ambiguous questions | 75 |
| | Unfair questions | 78 |
| | Unbalanced line of inquiry | 79 |
| 7 | WRITING CLEAR QUESTIONS | 80 |
| | Reduce the sentence length | 80 |
| | Simplify the word structure | 80 |
| | Be careful about words with several meanings and other problem words | 81 |
| | Do not use abstract words | 81 |
| | Reduce the complexity of ideas and present them one at a time in logical order | 82 |
| | Simplify the sentence structure | 82 |
| | Use active and passive voice appropriately | 83 |
| | Use direct, periodic, and balanced styles appropriately | 83 |
| | Avoid writing styles that inhibit comprehension | 83 |
| 8 | DEVELOPING UNSCALED RESPONSE LISTS | 86 |
| | Developing comprehensive lists | 86 |
| | Presenting mutually exclusive categories | 87 |
| | Using relevant and appropriate categories | 88 |
| | Using categories of appropriate specificity | 89 |

| | | <u>Page</u> |
|---------|---|-------------|
| CHAPTER | | |
| | Listing categories in the logical order expected by respondents | 89 |
| | Keeping the response list reasonably short | 90 |
| | Using a screening question | 90 |
| 9 | MINIMIZING QUESTION BIAS AND MEMORY ERRORS | 91 |
| | Question bias | 91 |
| | Memory errors | 97 |
| 10 | MINIMIZING RESPONDENT BIASES | 101 |
| | Response styles | 101 |
| | Highly sensitive items | 104 |
| 11 | MEASUREMENT ERROR AND MEASUREMENT SCALES | 108 |
| | Measurement error | 108 |
| | Measurement scales | 110 |
| 12 | ORGANIZING THE LINE OF INQUIRY | 114 |
| | Setting expectations about our line of inquiry | 114 |
| | Sequencing questions | 114 |
| | Using subtitles as cues | 115 |
| | Choosing an opening question | 116 |
| | Obtaining complex data | 117 |
| | Using transitional phrases | 118 |
| | Putting specific questions before overall judgment questions | 118 |
| | Dealing with adverse interactions | 119 |
| | Anticipating respondents' reactions | 120 |
| 13 | QUALITY-ASSURANCE PROCEDURES | 121 |
| | Pretesting | 122 |
| | Expert review | 128 |
| | Verification | 129 |
| | Validation | 129 |
| | Analysis of questionnaire nonresponses | 131 |
| | Testing reliability | 133 |
| 14 | FORM DESIGN AND LAYOUT | 135 |
| | Instructions | 135 |
| | Form preparation | 136 |
| | The style of the form | 136 |
| 15 | PREPARING THE MAIL-OUT PACKAGE AND COLLECTING AND REDUCING THE DATA | 143 |
| | Preparation of the mail-out package | 143 |
| | Data collection | 147 |
| | Data reduction | 150 |

| | | <u>Page</u> |
|----------|--|-------------|
| CHAPTER | | |
| 16 | ANALYSIS OF QUESTIONNAIRE RESULTS | 152 |
| | Analysis plan | 152 |
| | Item responses and univariate analysis | 153 |
| | Bivariate analysis and comparison of two groups | 153 |
| | Multivariate analysis and comparison of multiple groups | 153 |
| | Choice of analysis methods | 154 |
| APPENDIX | | |
| I | Bibliography | 155 |
| FIGURE | | |
| 1.1 | Typical completion times for major questionnaire tasks | 15 |
| 12.1 | Sequence of questions on funding data | 117 |
| GLOSSARY | | 156 |

CHAPTER 1

WHAT WE NEED TO KNOW ABOUT QUESTIONNAIRES:

AN OVERVIEW

THE ROLE OF QUESTIONNAIRES IN GAO EVALUATIONS

To meet GAO's broad mandate to study and evaluate federal programs, services, and funding, we draw from a wide range of evaluation methods. This transfer paper focuses on the use of questionnaires, which constitute an important data collection technique for implementing evaluations.¹

Most GAO work is directed at answering project questions that may be descriptive, normative, or cause-and-effect. By descriptive questions, we mean questions that ask about the condition of the entity under study. Normative questions ask how well the observed results of a program compare with a norm, criterion, or expected level of performance. A cause-and-effect question asks whether a program or policy caused particular outcomes.² Questionnaires may be part of the data collection method for answering each of these three broad types of project question.

To answer evaluation questions, four broad strategies are most commonly employed: sample surveys, case studies, field experiments, and available data. In sample surveys, data are collected from a sample of a universe to determine the universe characteristics, such as their range or dispersion, the frequency of occurrence of events, or the expected values of important universe parameters. A case study is an analytic description of the properties, processes, conditions, or variable relationships of either single or multiple units under study. A field experiment seeks the answer to a cause-and-effect question by contrasting the outcomes associated with a program to an estimate of what the outcomes would have been in the absence of the program. Use of available data as a strategy refers to the analysis of data previously collected or available from other sources such as the current population survey.

Sample survey and case study strategies are usually used to answer descriptive and normative questions. Field experiments address cause-and-effect questions. Depending on the situation, available data strategies can be used to answer all three types

¹We use the term "evaluation" throughout this paper, but its concepts and procedures apply equally to many GAO audits.

²These concepts are discussed more fully in Designing Evaluations, PEMD methodology transfer paper 4.

of question: descriptive, normative, and cause-and-effect. Original data collection strategies such as mail questionnaires are often used in sample surveys, but they may also be used in case studies and field experiments.

During the planning phase, the use of an evaluation strategy that requires the collection of original data must involve a consideration of the kind of information to be acquired, the source of information, the method for collecting data, the timing and frequency of the data collection, the sampling strategy, and the data analysis plan. For the field experiment strategy, we need also to consider the comparison base that will be used in drawing conclusions about cause and effect.

To see how the choice of a data collection method fits into the planning process, suppose that we are seeking information about the use of services from a federal program. We may get this information from records, from people who use the program services, or by making observations as services are provided. If we settle on collecting data from service users, then self-administered questionnaires might be one of the methods used to collect data.

We qualify the possibility of using self-administered questionnaires for two reasons. First, the choice of the data collection method, in this case self-administered questionnaires, must be compatible with the other design elements (such as the timing and frequency of data collection, the sampling strategy, and the analysis plan). Second, there are several techniques for collecting these data; examples are field observations of usage, usage records, personnel or telephone interviews (structured or unstructured), and self-administered questionnaires (structured or unstructured).³ The self-administered or mail questionnaire is thus one among many data collection techniques; it should be selected only if it is considered to be the most appropriate of these techniques.

Hence, the decision to use a self-administered questionnaire should be based not on an arbitrary choice but, rather, on a careful consideration of the evaluation question, the strategies for answering the question, and the other elements of the design.

WHAT WE HAVE TO KNOW TO USE QUESTIONNAIRES

The use of mail or self-administered questionnaires is an important and popular technique for data collection.⁴ It is

³For more detailed discussion on structured interview techniques, the reader is referred to Using Structured Interviewing Techniques, PEMD methodology transfer paper 5.

⁴Most self-administered questionnaires are mail questionnaires.

popular because much of our work requires gathering information from special populations of people who have firsthand knowledge and experience and because it is usually more cost effective than other comparable techniques, such as the personal interview, for gathering expert information.

But developing and using high-quality questionnaires--those that provide the information we need at the lowest cost--is never an easy job. For large-scale, complex evaluations, the job is even more difficult. And many GAO projects are complex; the typical job that uses a questionnaire involves the measurement of about 170 conditions or variables.

Some believe that most program specialists can easily write questionnaires without special training. However, writing good questions is like most other audit and evaluation tasks; to do a good job, we have to learn and work at our trade. We have to learn specific specialized skills. The same is true for designing self-administered questionnaires.

Writing questionnaires is the science and art of asking the "right" questions of the "right" people in the "right" way. It is a science in that it uses many scientific principles developed from various fields of applied psychology, sociology, 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 and understand. We have to learn how to write questions in a clear, concise, interesting, and easy-to-read format with a minimum loss in qualifying precision.

To run an evaluation at GAO that uses a mail or self-administered questionnaire, it is useful to know (1) GAO's approach to questionnaires (discussed later in this chapter), (2) when and when not to use a questionnaire (discussed in detail in chapter 2), (3) the questions that should be included in the questionnaire (discussed in chapter 3), (4) some of the basic principles for writing and organizing questions (chapters 5-12), and (5) the tasks involved in developing and using questionnaires (chapters 4 and 13-16). These topics are covered in brief in the remainder of this chapter.

GAO's approach to questionnaires

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

Our questionnaires 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.

Sometimes we are required by law to ask for specific information. However, our questionnaires usually cover many topics, and it is rare for all items to be mandatory. Hence, we usually make our requests voluntary.

We usually approach respondents by name with a polite appeal for their help. We use a signed transmittal letter that explains what GAO is, the purpose of the questionnaire, and how and why the respondents were selected. When it is appropriate, we also provide assurances of confidentiality and anonymity. We sometimes use questionnaires for investigative purposes but we rarely use them to identify specific persons for enforcement action. When we do, we clearly state this purpose and emphasize that the person has the right to refuse to answer.

Generally, our questionnaires are highly structured; they pose a collection of questions in a standardized and precise fashion. "Standardized" means every recipient is asked the same question in the same way. "Precise" means the questions are asked as exactly as possible. Because the questionnaires are self-administered, they must clearly communicate what the questioner needs to know and must provide a way for the respondent to give unambiguous answers.

An unstructured, imprecise item might ask, "What kind of work do you do?" A current homemaker might list "teacher" as an occupation held many years ago; a newly hired bank employee with minor administrative tasks might list "manager." To provide structure and precision to this question, we might ask several questions: "Which occupational category best describes the kind of work you currently do? What firm or organization are you currently employed by? What are your major job tasks?" Notice that these questions specify a time. In addition, because occupational categories may mean different things to different people, respondents are asked to describe their job tasks. To further structure the question and to help interpret the responses, we could provide a structured response format showing a complete range of occupational categories, including examples. The following example lists some of the structured responses available:

1. ☐ Clerical or kindred worker, such as bank teller, cashier, postal clerk, dispatcher, etc.

2. ☐ Technical worker, such as drafter, computer operator, laboratory technician, photographer, etc.
3. ☐ Professional worker, manager, or administrator, such as nurse, engineer, teacher, computer programmer, business administrator, insurance sales person, etc.

Unstructured questions are usually broad in scope and permit the respondent to give a totally free answer. They are generally most useful during the early stage of an evaluation when exploring a problem. Sometimes they are our best option, either because we do not have enough knowledge to adequately structure the question or because we are concerned that the structure might unduly influence the respondents.

However, unstructured questions can seriously compromise the validity of conclusions reached by means of mail or self-administered questionnaires, because they decrease the likelihood of obtaining relevant, comprehensive, and unbiased data. They may be variously interpreted; respondents may tell only what they remember at the moment and the reasons for omissions are not always obvious. Furthermore, analysis, aggregation, and presentation of the data later in the evaluation may become very difficult because of the questionnaire's lack of structure and the incomparability of the responses.

In contrast, specifically phrased questions with appropriate response choices reduce the likelihood of obtaining partial or wrong answers and improve the ease of data analysis, because they provide a framework and cues and because they are structured with the data analysis in mind.

When and when not to use mail questionnaires

Mail questionnaires are useful when we need 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 and it is difficult or costly to contact them by telephone or personal visit. Questionnaires are difficult to use if the respondent population cannot be readily identified by name and address, if the population is not easily reached by mail, or if the information we are seeking is not widely distributed among the population of those who hold the knowledge. Furthermore, questionnaires should not be used if the population does not have the required level of literacy or if the respondents are unable or unwilling to provide accurate and unbiased answers. Questionnaires are also not as effective as other techniques if the data collection time is very short.

What questions should be asked of which respondents

Because it is difficult to write good questionnaires, they are one of the most misused of data collection techniques. A few decades ago, a poll of the leading practitioners noted that poor question wording, sampling, and interpretation of results were the leading problems in survey research. We doubt that much has changed. A few years ago, GAO reviewed a random sample of 300 surveys used by the federal government. Most of the surveys were flawed by poor questionnaire construction as well as by inadequacies of study design, instrument testing, sampling, or implementation procedures.

Part of the problem is that good practices cannot be well documented by a few easy-to-remember principles. For every rule on questionnaire design and implementation, there is a host of exceptions. Another problem is competing requirements; the final questions are usually a trade-off between many principles that work against one another. Another barrier is that the field is not well documented with handbooks or tutorial texts. It is rare to find a university that offers a course in questionnaire design, and in the last 30 years, only a dozen or so textbooks have addressed the development and use of questionnaires. (See appendix I for a bibliography.) And because the field is so broad, these books treat only some of the fundamentals. The state of the art remains buried in thousands of journals, such as the Public Opinion Quarterly, and in the trial-and-error experiences of seasoned practitioners.

Another problem is that the development and use of questionnaires looks easy but is not. How do we determine the right questions to include in the data collection instrument? First, we must carefully analyze the overall questions the project was designed to answer. The line of questioning on the instrument, along with other aspects of the evaluation design, must lead to answers to the project questions. Second, the questions in the instrument must be asked in a way such that different people with very different experiences will provide similar answers under similar circumstances.

How do we find the "right people" when those we need are experts with firsthand knowledge in electrically charged particle beams or Apache Indians or migrant workers? We cannot just contact a few people who are conveniently located or appear to be approachable. We have to find out if they can and will give us the information we need. Then we have to select samples of respondents that are compatible with our overall evaluation strategy. (See Using Statistical Sampling for detailed information on drawing samples in ways that permit inferences to the sampled universe.)

Asking good questions is very hard to do. They must be asked in a way that encourages people to respond--and to respond

accurately. Asking questions triggers a very complex and not very well understood introspective and cognitive process. Respondents have to understand what is being asked, retrieve relevant information from memory, analyze this information, make judgments about which information best answers the question, perhaps combine this information, and select an answer. People are different; they read, perceive, think, interpret, value, remember, and respond differently. Unless we are knowledgeable about, and can anticipate, their cognitive processes and adjust our inquiry to account for these differences, we may get as many different answers to the same question as there are people.

Some fundamental principles for developing questions

The eight basic principles for writing good questions are listed below and are discussed in detail in the chapters indicated. Underlying these rules is a most important axiom: We must be thoroughly familiar with the respondent group and must understand the subject matter from its members' points of view.

1. Ask questions in a format that is appropriate to the questions' purpose and the information required. (See chapter 5.)
2. Make sure the questions are relevant, proper, and qualified as needed. (See chapter 6.)
3. Write clear, concise questions at the respondent's language level. (See chapter 7.)
4. Give the respondent a chance to answer by providing a comprehensive list of relevant, mutually exclusive responses from which to choose. (See chapter 8.)
5. Ask nonbiased questions by using appropriate formats and item constructions and by presenting all important factors in the proper sequence. (See chapter 9.)
6. Get nonbiased answers by anticipating and accounting for various respondent tendencies. (See chapter 10.)
7. Quantify the response measures where possible. (See chapter 11.)
8. Provide a logical and unbiased line of inquiry to keep the reader's attention and make the response task easier. (See chapter 12.)

Tasks involved in developing and using questions

In the development and use of a questionnaire, regardless of the project's complexity, very few of the major tasks can be

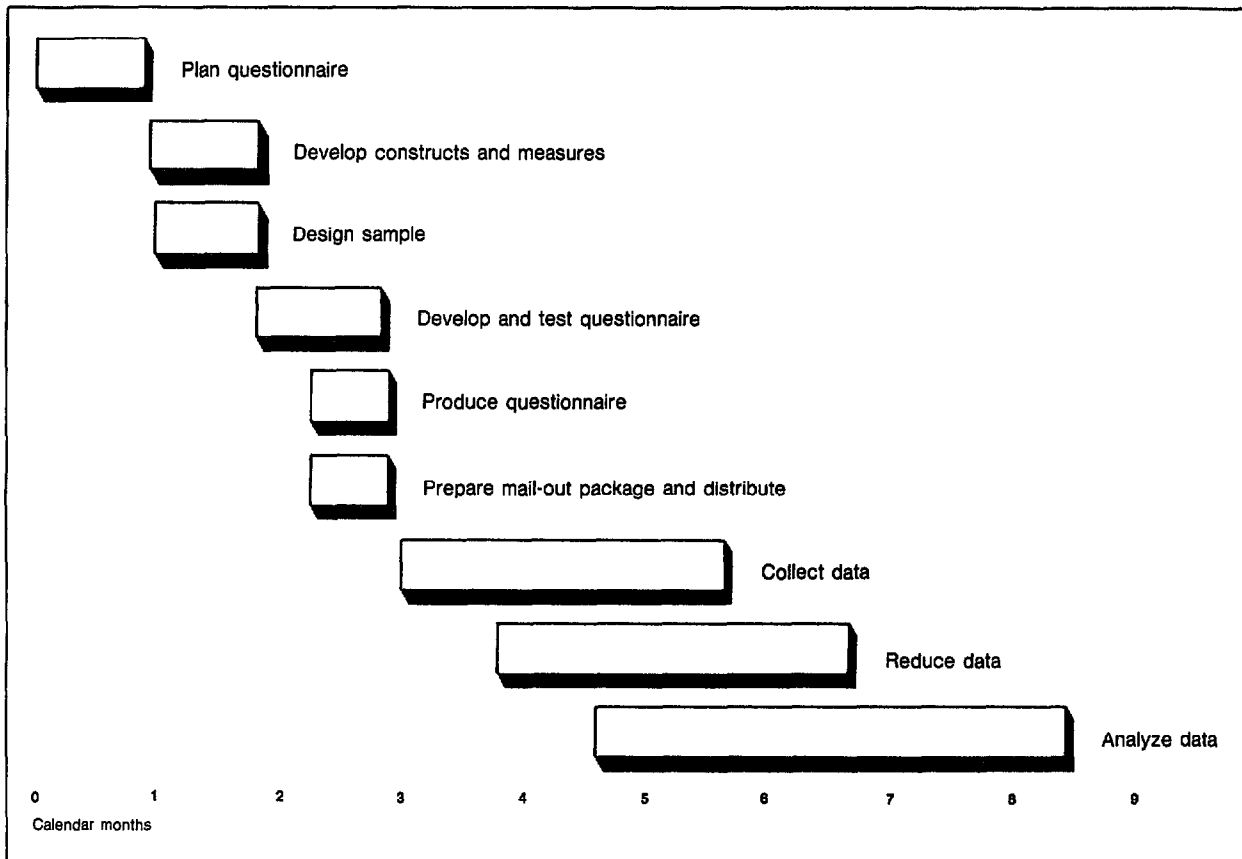
completed in parallel. Certain major tasks must be completed in logical sequence before others can be started. We must design the evaluation before beginning questionnaire development. We must decide what to measure before we can measure it. We must develop the questionnaire before we can test and validate it. We must design and print the mailing materials and address files before we send the questionnaires out. We must collect the data before we can prepare them for computer processing. And we must analyze and interpret the results before we can write the report. Once the decision is made to use a self-administered questionnaire as part of the evaluation design, the sequence of major tasks is

1. initial planning of the questionnaire,
2. developing the measures,
3. designing the sample,
4. developing and testing the questionnaire,
5. producing the questionnaire,
6. preparing and distributing mailing materials,
7. collecting data,
8. reducing the data to forms that can be analyzed, and
9. analyzing the data.

Although these tasks are not as labor intensive as those required for some other data collection methods, they often require considerable calendar time. The typical GAO mail survey takes about 9 months from planning the questionnaire through reporting the results. Simple data collection by questionnaires can be completed in about 6 months, but completing very complex, controversial, or otherwise difficult efforts might take 18 months. The amount of time depends not only on the evaluation's complexity but also on the availability of a satisfactory list of respondents, including addresses, and the number of different questionnaires to be developed for the project. Adequate time spent up front on evaluation design and planning and consensus from all staff about the final questionnaire are key factors in doing the job in the shortest time possible. Extensive redesign of questionnaires because of a change in evaluation objectives or other changes can translate into a loss of many weeks. This is not to say that all design changes can always be avoided; however, all changes that can be avoided should be, by thinking out the job design in the finest detail possible before beginning the questionnaire.

The time required for many questionnaire tasks cannot easily be shortened. For example, pretesting and pretest revisions

Figure 1.1: Typical Completion Times for Major Questionnaire Tasks



ordinarily take at least a month; commercial typesetting takes at least 1 week. Key punching mailing address lists takes 3 to 7 days, depending on the clarity and length of the lists and the contractor's work load. The biggest time constraint, of course, is response time to both the initial mailings and follow-ups. To achieve the desired response rate of at least 75 percent for most applications, it is usually necessary to make one or two follow-ups to the initial mailing. (See chapter 16.) In some instances, the data collection time can be shortened by using telephone or mailgram follow-ups, but these are expensive. Figure 1.1 shows time lines for a typical GAO mail questionnaire. The list on pages 16 and 17 outlines the specific tasks to be accomplished from the moment the decision has been made to design a questionnaire.

Tasks in developing and using questionnaires

1. Initial planning of the questionnaire

Review the evaluation design (p. 29)

Specify the variables (p. 30)

2. Developing the measures

Operationalize the variables (pp. 30-32)

Identify standards for comparing the data (pp. 33-34)

Relate measures to units of analysis and target populations (pp. 34-35)

Develop plan for data analysis, validation, and statistical approaches (pp. 35-36)

3. Designing the sample

When a questionnaire is applied to the entire universe, it is not a sample but a census. If no universe can be identified, you may need a judgment sample.

Specify universe, evaluate its adequacy and efficiency, and identify and assess its source (pp. 37-41)

Decide on a statistical or nonstatistical sample (pp. 44-46)

Develop sampling design, adequacy and efficiency, and develop sample selection procedures (pp. 41-44)

4. Developing and testing the questionnaire

Write the questions (chs. 5-11)

Organize the questions into a line of inquiry (ch. 12)

Design the pretest form (p. 136)

Develop pretest procedures, select pretest sites and units, and train interviewers (pp. 121-25)

Conduct pretests (pp. 125-28)

Obtain expert reviews and peer reviews (pp. 128-29)

Analyze pretests and expert peer comments (pp. 121-29)

Revise pretest form for final instrument design (pp. 135-42)

Develop and test validation instruments (pp. 135-36 and 129-31)

5. Producing the questionnaire

Design form and produce camera-ready or illustration-ready copy; supervise and coordinate printing and illustration services (pp. 135-42)

6. Preparing and mailing materials

Develop address lists and edit for keypunching onto computer tape or other device (pp. 143-44)

Develop transmittal letter (pp. 144-47)

Duplicate transmittal letter and enter address list into appropriate device (p. 147)

Design and assemble mail-out materials (p. 147)

Check and correct mailing lists (p. 144)

Distribute materials (pp. 147-48)

7. Collecting data

Specify follow-up, nondeliverable, and mortality analysis procedures and execute them (pp. 147-50)

Log and track returns (p. 148)

Specify and supervise data collection procedures (pp. 147-50)

8. Reducing the data to forms that can be analyzed

Edit returned questionnaires for response consistency and adequacy of keypunch entries (pp. 148-50)

Manage data base loading into computer (p. 151)

Specify keypunch verification procedures and make verifications (pp. 150-51)

Clean up data base and check item responses (pp. 150-51)

9. Analyzing the data

Develop codebooks (pp. 152-53)

Produce descriptive statistics (p. 153)

Update analysis plan (pp. 152-53)

Conduct multivariate analysis when appropriate (pp. 153-54)

Interpret analysis and draw conclusions (pp. 153-54)

CHAPTER 2

ADVANTAGES AND DISADVANTAGES

OF MAIL QUESTIONNAIRES

CHOOSING THE DATA COLLECTION METHOD

We have noted that data can be collected in a variety of ways, such as personal interviews, mail questionnaires, telephone surveys, and reviews of records. Each method has strengths and weaknesses. The project objectives, the evaluation strategy (such as a sample survey, a case study, or a field experiment), and the data sources must be considered in choosing data collection methods. Also, 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. Questionnaires are frequently used with sample survey strategies to answer descriptive and normative audit or evaluation questions. They are typically less central in studies answering cause-and-effect questions with nonequivalent comparison group or interrupted time-series designs.

The project objectives, the evaluation strategy, and practical constraints thus influence the choice of data collection method. For example, if the objectives of the project were to determine the average per acre charge and the income derived from public grazing-land permit fees, the evaluation designer might consider using auditors or clerks and structured manual data collection forms or pro forma work papers to survey the case files in record storage. However, if the objectives were to determine how much land and how much per acre the ranchers were willing to lease or pay, then a mail or telephone survey of ranchers would be necessary.

Even if the research strategy calls for a sample survey, a prime situation for questionnaires, the decision to use mail questionnaires should be made only after a careful consideration of the guidelines presented in this chapter. Although generally questionnaires are very versatile, there are many situations in which other data collection techniques may be superior. This is not a rare event; indeed, other techniques were recommended for about one of every three proposed GAO questionnaires discussed with PEMD in 1985.

Before choosing a questionnaire over other data collection techniques, it is important to consider both the advantages and disadvantages of mail questionnaires compared to other methods of data collection, such as telephone and personal interviews, review of records, and the use of extant data and field observations. Some advantages and disadvantages (which, in most instances, apply also to self-administered questionnaires) are summarized in the following list.

Advantages

1. Mail questionnaires are generally more versatile than other methods (p. 20)
2. Like interview techniques, mail questionnaires are usually more compatible with survey designs than other methods (p. 21)
3. Except for the use of certain types of extant data, mail questionnaires usually cost less than other methods (p. 21)
4. As with extant data and field observations, mail questionnaires have no interviewer bias (p. 22)
5. Mail questionnaires have less response bias from social desirability and question threat than interview methods (p. 22)
6. Mail questionnaires usually permit a wider distribution of the sample, if sampling is used, than other methods (p. 22)
7. With certain exceptions, mail questionnaires usually provide easier access to the data sources than other methods. The exceptions apply to specific instances of random digit dialing and extant data collections (p. 22)¹
8. Save for field observations, mail questionnaires usually afford greater opportunity to collect detailed data that cannot be immediately recalled, check records, or consult colleagues and other sources (p. 23)

Disadvantages

1. More uncertainty as to respondents' identity with mail questionnaires than with other methods (p. 24)
2. Mail questionnaires often require longer turnaround times than other data collection techniques (p. 24)
3. Nonresponse can be a problem more with mail questionnaires than other techniques (p. 24)
4. When the identification and location of knowledgeable respondents is difficult, other data collection techniques are usually more appropriate than mail questionnaires (p. 24)

¹Random-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.

5. Mail questionnaires are more difficult to apply than other techniques if literacy among the respondents is low (p. 25)
6. Both interview and mail questionnaire methods may give more-distorted results than other methods when there is a possibility of biased reporting (p. 25)
7. Both interview and mail questionnaire methods are less suited than other techniques if the respondents are likely to be inaccurate reporters (p. 25)
8. Interviews, extant data, and field observations usually have an advantage over mail questionnaires if the information is not uniformly available from respondents (p. 26)
9. Both mail questionnaires and the collection of extant data are not as well suited as interview and field observation methods to exploratory, broad-based, or complicated methods of inquiry (p. 26)
10. Mail questionnaires are harder to implement than other methods if the nonresponse is focused or concentrated (p. 28)

COMPARATIVE ADVANTAGES OF MAIL QUESTIONNAIRES

Versatility

Mail questionnaires are more versatile than face-to-face and telephone interviews and extant-data and field observation data collection techniques. By this, we mean that they can be more readily used both to collect many types of information and to collect information from a wider variety of sources than other methods. Mail questionnaires can be designed to gather a variety of different kinds of information (for example, facts, figures, amounts, statistics, dates, attitudes, opinions, experiences, events, assessments, and judgments) from a single type of knowledge holder. Examples include descriptions and problems encountered in establishing professional standards review organizations (PSROs), costs and membership figures and satisfaction with the services of health maintenance organizations, and the current and projected amount of research and development costs due to the U.S. Treasury on foreign sales contracts as reported and estimated by munitions control license holders. Other data collection methods are not as well suited as mail questionnaires for extensive and detailed data collection of different kinds of information.

Mail questionnaires can also be used to collect data from far-ranging and different sources. For instance, one study may use variations in a questionnaire to collect similar data from local, state, and federal officials. Another might record organizational effectiveness data from production and personnel records as well as from maintenance staff, blue collar workers, clerical workers, support staff, professional staff, and line,

middle, and upper management personnel. The respondents' inconvenience, privacy concerns, respondent memory requirements, and lack of flexibility constrain other methods in collecting these kinds of information more than they do mail questionnaires.

Compatibility with survey designs

Another advantage of mail questionnaires, as well as face-to-face and telephone interviews, is that they are very compatible with survey designs. This is an important consideration, because the sample survey is one of our most widely used design strategies. It can be employed by itself or in combination with other strategies to

- generalize findings to the universe under study,
- describe and occasionally explain phenomena, and
- complement, crossvalidate, or verify another evaluation strategy, such as the case study.

For example, in a report on Pell grants, GAO assessed the effect on students of requiring them to prove their eligibility.² The assessment was based on 12 before-and-after case studies, and the authors could not generalize without additional evidence. The ability to generalize was provided by a questionnaire sent to 400 randomly selected schools. The results from the national survey corroborated the findings developed from the more rigorous study of the limited number of cases. It is usually much more difficult to use extant data or field observations on a large scale for these purposes than it is to use questionnaires or interviews.

Low cost

Except for certain uses of already existing data, questionnaires are often less expensive than other types of original data collection techniques. For example, a mail questionnaire response can be obtained for \$1 or \$2 a case, whereas telephone and face-to-face surveys have respective cost ranges of \$25 to \$50 and \$100 to \$300 per case. Data collection by field and record observations may provide information comparable to information from a mail survey but usually at much greater cost. Questionnaires are inexpensive because there are no interviewers or observers and, hence, no associated salary, recruiting, or training costs and because follow-up costs are low.

²Pell Grant Validation Imposes Some Costs and Does Not Greatly Reduce Award Errors: New Strategies Are Needed, GAO/PEMD-85-10, September 27, 1985.

No interviewer bias

When a personal or telephone interviewer varies the way a question is asked or when the interviewers differ among themselves in the way they ask the questions, bias and variability are introduced. For a well-controlled study, this can account for 10 or 20 percent of the error, and studies that are not well controlled may be invalid. Since there are no interviewers in a mail survey, this source of error is avoided.

Reduced response bias

Research on data collection techniques shows that mail questionnaires usually reduce respondent bias. Respondents are more likely to answer difficult, unpleasant, or threatening questions honestly if the questions are posed on paper rather than in person.

Wider distribution of sample

Mail questionnaires provide easy access to a large number of data sources. With names and addresses, statistical sampling that permits generalization to large populations is feasible, regardless of distance, language, and access barriers. For instance, in several studies we were able to collect data with relatively little cost and effort from difficult-to-contact scientists, administrators, and other knowledge holders, who spoke different languages and were located in different parts of the world. This would have been extremely difficult with other data collection methods such as personal and telephone interviews.

Ease of contact

One of the problems with personal and telephone interviews for large samples is that each person is likely to have a different time and location convenient for an interview. Hence, without complex scheduling and prior appointments, which are also difficult and costly to obtain, it is not unusual to make several telephone calls at varying times on different days. Personal interviews also require similarly onerous call-backs. Field-observation data-source contacts require careful scheduling. Using existing data may also pose data source contact problems, depending on the circumstances.

Hence, for large samples, data collection techniques other than the mail questionnaire are often not feasible. This becomes readily apparent when we compare data collection efforts on a typical job requiring contact with 1,000 persons (exclusive of travel): 5 or 6 staff days by mail questionnaire, 175 staff days by telephone, and 250 to 1,000 staff days by personal interview.

Ability to collect detailed data

If the respondents are literate and do not have to compose a written answer, they can usually process more complex information faster in a printed mode than in an oral mode. This is because a printed format usually makes information available when it is needed and substantially reduces the immediate or short-term memory burden. For example, in an interview, a respondent may have to commit to memory the seven choices from which to make a selection. In a mail questionnaire, the seven choices are immediately perceived by visual inspection.

Reading written information and making answer check marks are usually much faster than understanding and responding to oral communication. On a typical mail data collection instrument, a respondent can provide answers to 50 questions requiring 14 pages of text in half an hour. The same instrument would require an hour and a half in a personal interview and more if administered by phone. Since the respondent burden is often considerably less with the written form, evaluators can usually collect more information and more detailed information with mail questionnaires than with other formats.

Respondents also have more time to retrieve detailed information with mail questionnaires than with other methods. They have time to recall data that had been temporarily forgotten, check records, send for information, and consult with colleagues and other sources. Interview and field observation methods are more immediate and do not afford respondents this opportunity.

COMPARATIVE DISADVANTAGES OF MAIL QUESTIONNAIRES

Uncertainty of respondent identity

We should not send a mail questionnaire unless we are sure that it will be completed by the intended respondent. Questionnaires sent in packets to be distributed around the office or work site that do not identify specific people can sometimes be completed by someone other than the intended respondent. Similarly, a questionnaire sent to a specific address with a broad title such as manager, administrator, staff, training officer, resident, or student may be filled out by someone other than the one intended because there are many kinds of administrators, managers, staff, and so on. With telephone or personal interviews and other methods, data collectors have a chance to test the eligibility of the respondent or their sources and screen out inappropriate selections. Not so for mail questionnaires. Hence, if we are not sure that we have the appropriate respondent, we should not use mail questionnaires, because conclusions drawn from the responses would be subject to challenge.

Long turnaround time

As we already noted, mail questionnaires typically take 9 months from beginning to end (from questionnaire design to results); that is, they take longer to design than other instruments. There are two reasons for this. First, because an interviewer is not present to help the respondent through any difficulties, the questionnaire must be designed carefully so that no difficulties can arise.

Second, mail questionnaires often require many items, because they are used in complex projects, whereas some other data collection methods, such as telephone and personal interviews, are more applicable to simpler data collection. The typical GAO mail questionnaire contains 12 pages, 34 questions, and 167 variables. This complexity adds considerably to the design time.

Consistent with the task-completion time discussed in chapter 1, it may take about 3 months to plan, draft, test, evaluate, revise, and print a questionnaire. It then takes 2 or 3 months to mail out and receive responses to an initial request and conduct two follow-up appeals. Data processing, analysis, and interpretation usually take another 3 months. When evaluations have really tight deadlines, complex mail questionnaires may not be the method of choice.

Possible nonresponse problems

In personal or telephone interviews, we can keep trying until we contact the respondents. Once we have made contact, they are less likely to turn us down. Similarly, extant-data and field-observation methods are usually designed to keep the number of missing observations low. But in mail questionnaires, if we have not heard from respondents after three or four follow-ups, they are not likely to answer. Because of this, only proven questionnaires or other alternatives with high-yield data-collection techniques should be used when we expect to encounter a nonresponsive population.

Identification and location of knowledgeable respondents

A mail questionnaire is not appropriate unless we can readily identify the respondents and get their mailing addresses in advance of data collection. For example, the informants on illegal aliens and other "hidden populations" often have to be located by "snowball" sampling. That is, during data collection, each informant identifies the next informant until no new informants are identified. This type of sampling is difficult by mail. Similarly, we had to abandon the questionnaire approach in an evaluation of information resource management, because we could not locate within each of the agencies people sufficiently knowledgeable in the topic.

Some populations cause particular difficulty. Individual armed forces recruits are difficult to locate because they move around a great deal during the first few months of training. Migrant workers have no fixed address during a harvesting season.

Data sources may also be hidden in a much larger source. For example, the names of major importers and exporters are embedded in a list of 10,000 importers and exporters. Among the 10,000, most of the businesses are too small to be of interest. Unless we find a way to locate the small group of importers and exporters who account for the major portion of business, a mail questionnaire is not feasible. Interview techniques, not mail questionnaires, usually have the best chance of success when the identification or location of knowledge holders is a problem.

Low literacy or poor vision among the respondents

Respondents must be able to read and follow a questionnaire. When substantial portions of a population read below the fourth-grade level, it is very difficult to develop a line of inquiry. Higher levels of literacy may be required if the questions involve difficult concepts, operations, or procedures. Also, mail questionnaires may not be appropriate or may need special modification for populations who tend to have poor vision, such as the elderly. Interview and other techniques usually have an advantage over mail questionnaires in situations like this.

Possible problem of bias

Topics that might prove threatening to respondents or cause them discomfort are inappropriate for both interviews and mail questionnaires. Questions requiring self-evaluation on alcohol and drug use, for example, produce very inaccurate responses unless special procedures are used. One questionnaire asked employees how many times they were late for work last month. It is not surprising that few admitted to having been late even once. In the few instances when they were late, they said they made up the time. This data collection had little credibility. A rewording of the question or another evaluation approach, such as a review of records or personal observation, would have produced better data.

Likely inaccuracies among respondents

Sometimes people are not knowledgeable or accurate reporters of certain kinds of information. They report some things very well and other things very poorly. For example, veterans might accurately report that doctors made medical examinations for agent orange on their eyes, ears, nose, throat, genitals, and rectum but might substantially underreport skin examinations. If we needed the information on skin examinations, other sources, such as medical records, might be more useful.

Lack of uniformly available data

The information sought must be uniformly available from the intended respondents. When informed data sources vary from agency to agency or among geographic locations (that is, when subgroups of the population holding the information we need differ), mail questionnaires may not be effective. For example, questionnaires may not be the best way to get data on how cities handle the problem of teenage prostitution, because cities use a variety of groups (police, charities, court systems, welfare agencies) to do this and the groups interact differently from city to city. For these and similar problems, interview and field observation techniques are likely to yield better results.

Exploratory, broad based, and complicated methods of inquiry

Mail questionnaires are not as well suited to broad exploratory work as other techniques, because they do not permit an exchange of information. Before developing a mail questionnaire, we must have enough information to understand the phenomenon being examined. We typically need a thorough knowledge of statutory history, programs, processes, conditions, outcomes, and criterion performance for the area that is under investigation. To develop a block grant questionnaire, for example, we needed to understand the 38 types of local delivery systems; the responsible agencies or offices; the major planning, implementation, and evaluation processes; and the major programs, activities, issues, and financing systems. We must already know what to measure before we begin writing a mail questionnaire. If we do not have this information, we must use other techniques, because we cannot use probes or feedback to get the knowledge holder to help us.

Broad and global questions are also not appropriate for mail questionnaires. If we ask such questions of a large number of respondents, we will not be able to accurately compare the responses. Answers will vary considerably, because the respondents do not know how to answer or what is expected of them. They have different frames of reference, and we cannot use probes or feedback to help focus on the right frame of reference at the appropriate level of detail. As a result, the analyst will not know whether the variability reflects a true state of affairs or measurement error--that is, in this case, respondent uncertainty or misunderstanding. Such questions as "What are the roles and responsibilities of program managers in major weapon systems?" should not be asked unless they can be broken down into specific elements or subquestions. If they cannot, more scoping or another data collection method (such as personal interviews with experts) may be called for.

Mail questionnaires are difficult to design if the method of inquiry is complicated by the use of multiple screens or

branching, complex prompting, or detailed instructions. The multiple branching instructions drafted for a survey of taxpayers who were receiving IRS assistance illustrate this kind of difficulty:

IMPORTANT: Your answer to Question 4 determines which other questions you will complete in the rest of this survey.

IF YOU CHECKED 4a,
please answer questions 5-7 and 18-26.

IF YOU CHECKED 4b,
please answer questions 8-10 and 18-26.

IF YOU CHECKED 4c,d,e, or f,
please answer questions 11-15 and 18-26.

IF YOU CHECKED 4g,
please answer questions 16-26.

We can see by inspection that if the respondents are not very careful, they will either miss their screen or confound the results by answering inappropriately.

Some studies need multiple and complex prompting. For instance, it has been shown that prompting helps victims of crime recall their experience. Hence, an interviewer's line of questioning may start with "Think about the times you went shopping in the last year." The interviewer may then proceed to ask a series of questions about a victim's experiences of crime while shopping. Then the line of questioning may start all over again with a new activity (for example, "Think about all the times you left your car parked last year"). New lines of questioning would be introduced until all activities of inquiry had been covered. We can readily see that this type of data collection is more easily done when feedback and live prompts are used.

Sometimes data collection instructions and procedures are complex. For example, in a survey of household nutrition, respondents were required to provide precise estimates of the size of food portions. Interviewers helped recall portion sizes by explaining the amounts (for example, ounces) associated with the various cups, spoons, glasses, bowls, plates, and other measures in the respondent's kitchen. In a survey of users of a complicated tax procedure, the interviewers (who were also accountants) sometimes had to help the respondents with calculations. In a survey of drug users, interviewers had to follow a special procedure that guaranteed the secrecy of the

respondents' answers.³ We do not mean to say that evaluators cannot use mail questionnaires for complex screens, prompts, or instructions. They can be used, and there may be no alternative. However, the design, implementation, and analysis will be much more difficult with a mail survey than with other techniques.

Expected nonresponse concentrated
in one segment of the sample

It is always important to investigate the characteristics of nonrespondents whenever possible, but nonresponse is usually not a serious problem if its rate is low and evenly distributed. However, nonresponse is a serious problem whenever the nonresponse rate is disproportionately large for a particular subgroup of people and when the attributes of this subgroup are important to the project. Another problem is that the statistical weighting procedures used by the analysis to account for this imbalance may artificially reduce the sample error. This can introduce an error in subsequent secondary analyses. It is usually easy to compensate for this error if both the sampling design and the analysis are simple. However, the adjustments can become quite laborious if the sample is highly stratified and the analysis is complex. At best, certain assumptions about the respondents must be made, and a great deal of extra work is needed to make these adjustments. At worst, the study may not be valid if the assumptions about the nonrespondents turn out to be incorrect. If it is feasible, avoid these problems by using high-response-rate techniques.

³It is called the "randomized response" procedure. It uses probabilities to conceal a respondent's answer. The method is so apparent that the subjects quickly realize that a response is and always will be secret.

CHAPTER 3

PLANNING THE QUESTIONNAIRE

AND DEVELOPING THE MEASURES

REVIEWING THE EVALUATION DESIGN AND SPECIFYING THE VARIABLES

Once the overall evaluation questions have been set and the design for answering the questions has been chosen, work on data collection procedures can begin. If the data collection involves the use of questionnaires, the instruments must be planned, developed, and administered. However, as we shall see, the design process is iterative and the development of plans for questionnaires may cause changes in the job design.

In planning the questionnaire, we must decide what variables we want to measure and how we want to measure them.¹ The job design poses the evaluation questions and provides the design necessary to answer these questions. A close review of the evaluation questions and the design helps us choose what to measure and how to measure and, indeed, forces a justification for each variable selected.

When we consider what to measure and how, we may confront limitations that force us to reconsider the evaluation questions or the design. For example, one initial study design was intended to show the effects of inner-city health-care clinics on the health of the urban poor. However, we found we could not measure health but we could measure the amount, and some aspects of the quality, of health-care services received. Thus, measurement considerations changed the overall project question from something that was impractical to something that was researchable.

After the proposed design has been set forth, it must be reviewed carefully to see if all the essential variables have been listed and can be measured with the necessary precision. For example, a project's objectives might be to describe area agencies on aging and to determine whether they are in place and performing as legislated. These objectives imply descriptive and normative questions. We might identify some relevant measures to describe an area agency on aging--for example, size, organizational structure and auspice, methods of operation, quality of staff, level of funding, performance, and, for the normative questions,

¹Variables are sometimes referred to as "constructs." In either case, we are thinking about something that we ultimately want to measure but, during the early planning stage, may be a little abstract. For example, one variable or construct might be "socioeconomic status," a concept that must be made more concrete before we can measure it.

standards for performance. However, a close review might show that we had left out specific important variables such as the community context and relationships with state offices.

For each variable, we must consider the degree of precision needed in the measurement and whether we can achieve it. For example, it might do little good to measure performance if our measures are so gross that we cannot tell whether the standards have been met.

The review should also consider whether the results can be generalized, if this is a goal of the project. For example, it may be of little value to measure indirect costs across organizations if the organizations do not all have the same standardized way of accounting for indirect costs. We could not generalize about indirect costs across the organizations studied.

The review should also help us identify all important cross-sectional and temporal relationships between the variables that we need to know about. Frequently, we want to know how two or more variables are related, cross-sectionally, across the units in our sample. For example, if we want to know if size or organizational structure are related to performance, we must collect the data so that we have performance measures by size and by structure. By "temporal relationships," we refer to the time or sequence in which measures are obtained. For example, if our intent is to measure change in the operation of organizations, we would like to be able to make measurements of the conditions over time.

Before we complete this initial phase (sometimes referred to as "initial planning") we should know what variables we need to measure, why we need to measure them, and how they relate to one another. For example, if the purpose of the measures is to assess the economic viability of a company, we would not just be concerned with profit margin or the percentage of sales that is the company's profit. This measure is incomplete because it does not show reinvestments and amount of surplus money. We really need to know if a company is generating enough return on investment so that it can afford to keep itself competitive by reinvesting and still produce enough profit or profit potential to be worth investing in. Hence, we need to measure not only the profit margin but also the amount of surplus money, after all expenditures and reinvestments, and the amount of reinvestment.

OPERATIONALIZING THE VARIABLES

So far, we have identified the things we want to measure only in broad terms called "variables" or "constructs." These are ideas about traits, properties, and characteristics, but they are not measures. To get measures, we must develop operational definitions that translate these variables into concrete things or events that we can count or ascribe a single

dimension to (such as the presence or absence of a trait or the extent to which the trait is present or absent). To do this, we first analyze the variable into its component parts or properties, then analyze the interrelationships of these properties, and finally assign a single measurement dimension to each property that denotes the degree to which the property is present or absent. For example, we may be concerned about a variable called "timeliness." We can analyze this variable into two parts: turnaround time and availability. We can see that these two component constructs are independent. We operationalize these two parts as specific traits, each with a single dimension: elapsed time between receipt of the request and delivery of the job product and whether or not the congressional committee got the product or project information in time to use it.

The illustration above shows a very simple variable to operationalize. The task is usually more complicated and requires a great deal of skill and innovation on the part of the question designer. The question writer has to become a knowledge broker, using various kinds of information sources to operationalize constructs or variables: textbooks, literature searches, site visits, reviews of legislation and other documents, reviews of other GAO audits, and interviews with key knowledge holders and spokespersons from both sides of the issue. For example, in a management study, we had to develop a measure for quantifying the typical manager's role in the Air Force systems command. Textbooks, literature reviews, past GAO studies, and the position descriptions of the command yielded several activities that could be operationalized to measure the manager's role: managing staff; conducting intergroup coordination; doing administrative work; performing functional and operational tasks other than management; performing liaison; conducting negotiations; performing spokesperson tasks; and planning, developing, organizing, evaluating, reporting, and developing new procedures and strategies for improvement or change. Each of these functions was further qualified as to job operations, duties, knowledge and skill requirements, and responsibilities. The functions were quantified as to the frequency and duration of occurrence.

Sometimes there are many choices of measures. For example, the variable "unemployment" may be operationally defined in at least four ways: (1) the number who are actively seeking work, (2) the number of married men seeking work, (3) the number seeking work who were formerly in the work force, and (4) the number seeking work plus the number who are not working and have stopped looking for work. The choice between these measures depends upon the evaluation questions being asked.

In summary, we must settle on operational definitions for the variables we want to measure by analyzing and identifying the component actions, events, behaviors, activities, or objects that are implicit in the variable of interest and by giving each of these a single dimension on which it can be scored. However,

having these measures still does not give us questionnaire items. To develop a questionnaire item like the one that follows

Consider the following factors or properties which determine the quality of the sensor imagery photographs. Rate the adequacy of the photographic images on each of these factors. Base your rating on your typical experience with this study. Skip this question if you do not use imagery photographs. (Be sure to check one and only one box in each row.)

| Factors/properties | More than adequate | Adequate | Marginal | Inadequate | Very inadequate | Do not use this factor |
|--|--------------------|----------|----------|------------|-----------------|------------------------|
| Spatial resolution | | | | | | |
| Picture sharpness, level of blur | | | | | | |
| Amount of cloud cover | | | | | | |
| Range of contrast (difference between lightest and darkest areas) | | | | | | |
| Levels of contrast (number of shades between lightest and darkest) | | | | | | |
| Background clutter (amount of nonrelevant information) | | | | | | |
| Range of colors | | | | | | |
| Levels of colors (number of shades of colors) | | | | | | |
| Level of image distortion | | | | | | |
| Graininess of the photographic print | | | | | | |
| Degree of synoptic coverage (given area covered) | | | | | | |
| Accuracy of the ground-distance scale | | | | | | |
| Registration or satellite drift | | | | | | |
| Accuracy and/or range of location gradient | | | | | | |
| Other (specify) | | | | | | |
| | | | | | | |
| | | | | | | |

We must put these measures into words for the respondent to read and answer. Considerations in writing items and questions are covered in chapters 5-12, but the example we have just seen illustrates the final step, the question writing, needed to develop an operationalized variable into a questionnaire item.

In a study on earth-orbiting satellite data, we needed to know what users thought about the adequacy of the photographic image. To get the questionnaire item, we operationalized the variable "adequacy of photographic image" by listing the properties of a photographic image and developing a scale for rating adequacy. Then we wrote a question that would get the user to rate these properties.

IDENTIFYING STANDARDS

For some projects, it is necessary to compare measures to "standard" values for the measures. For example, in a study of day-care centers, we might wish to compare the space for each child (in square feet per child) to a standard set for day-care centers. Or, in a compliance audit of a federal agency, we might judge the compliance of the agency's personnel system by comparing measures with standards on several different dimensions. Comparisons might be made separately on each dimension, or the measures might be combined in some fashion and compared to a grand standard. Evaluation and audit questions that imply such comparisons are called "normative questions" (see the 1984 PEMD transfer paper entitled Designing Evaluations).

The establishment of a standard requires a value judgment. Usually, the judgment is of a form that indicates a measured value is unacceptably high (or low) when the standard is exceeded (or not achieved). For example, standards are set by school districts for student-teacher ratios, by oil-exploration teams for the number of exploratory wells per line miles of seismic data, by health-care planners for infant mortality rates, and by keypunch firms for the number of errors per 1,000 strokes.

When we need a standard with which to answer a normative question, we can use three main approaches: (1) adopt a standard from an authoritative source, (2) ask experts in the substantive field to reach consensus on a standard, or (3) set the standard ourselves, basing it on a combination of empirical analysis and value judgment.

An authoritative source such as legislation, a regulation, or an administratively established program goal is frequently used as the basis for a standard. Set in this way, the standards predate the GAO evaluation.

When standards do not already exist, we may ask experts on the topic in question to arrive at a standard. For example, in a

study of disease control, we might ask epidemiologists and health planners to set a standard for a disease rate such that any observed values below the standard would constitute "satisfactory control." A variety of methods such as small group conferencing, the Delphi method, and the analytic hierarchy process can be used to systematically arrive at group consensus.

The third possibility for setting a standard would be GAO professional judgment, probably backed up by some empirical analysis. For example, we might compare students' achievement test scores in an experimental schools program against "standards" for comparable students in the population at large. The value judgment would come in setting the 75th percentile, for example, as the basis for comparison. The empirical analysis would come in computing the test score that corresponds to the 75th percentile in the general population. The standard would then be the test score at the 75th percentile.

Not all measures require standards, but when normative questions are posed, the identification of standards is an important evaluative step related to measurement. Systematic procedures should be used to establish defensible standards.

RELATING THE MEASURES TO THE UNIT OF ANALYSIS

While we are developing the list of variables and translating them into measures, we must be thinking ahead to sampling and data analysis. We must think about the units (people, groups of people, objects, and so on) we will want to focus on in our data analysis so that we can answer the overall evaluation questions. For example, if we had a question like "What services are offered by educational program X?" the answer might depend upon whether we collected and analyzed data by student, classroom, school, school district, or state. We have to settle on what is called the "unit of analysis."

Choosing the appropriate unit of analysis is sometimes complicated because we can often imagine several different universes (groups of students, classrooms, and schools) from which we could collect data. It is usually desirable to collect data from the same units that we want to use in our data analysis. We prefer not to mix units, because mixing them makes the results of data analysis more difficult to interpret. For example, we probably want to avoid collecting data from students on the educational services they are provided and then averaging across students in a classroom to get a classroom measure. It would be technically better to keep the data collection units the same as the analysis units. However, there are frequently trade-offs to be made, including considerations of feasibility and cost, and sometimes we do mix units. That is, we generalize to a universe different from that of our data collection units. For example, we would collect data on contractors and generalize to contracts.

But to do this we must have a statistically valid way of relating the data collection unit (contractors) to the data analysis unit (contracts).

Because of our concern about the units of analysis, it is important that we look ahead and use our plans for data analysis as a factor in choosing measures.

INITIAL ANALYSIS AND VALIDATION PLANS

During this phase, we should also develop initial analysis and validation plans. This is important because the project should not proceed until we have made these plans; we do not want to put other project tasks such as sampling, quality assurance, and analysis at risk. We say "initial," because we cannot complete these plans until we have drawn the sample, developed the measuring instruments, and collected the data.

The purpose of the initial analysis plan is to document the measures, the variable interrelationships of interest, the logic used to study these interrelationships, the units of analysis, the comparisons necessary to accomplish the objectives of the study, and the precision and level of certainty that we expect to result from the data analyses. The sampling experts need this plan before they can determine the sampling strategies and sample sizes. The analysts, in turn, need both the initial plan and the sample design before they can determine the types of analysis, the computer and analytic support, and the software packages needed to complete the job. Hence, the effort starts during the questionnaire planning but it is a continuing effort that is updated during the course of the study as more information becomes available. For instance, it is updated when the sample sizes have been determined, after the measures have been made final and coded, and after the data have been collected.

Changes may be made to handle unforeseen or untoward events and new discoveries, but most of the updating will be to include additional details. Despite this necessary evolution, the structure of the analysis plan will be apparent from the start, because it is determined by the evaluation questions and time and cost constraints. For example, if the study is designed to provide simple descriptive information, much of the data will be analyzed by single variable methods. However, if questions are posed about the relationships between variables, then bivariable or multivariable methods must be used. The choice and complexity of these methods will be determined by the scope, complexity, and logic of the evaluation design.

During this phase, we must also develop an initial validation plan. "Validation" refers to a process of ensuring that we are really measuring the variables or constructs we say are measuring. We must validate because we cannot always be sure that our measures reflect or describe the conditions or ideas that we

intended to study. We validate when we have concern about the quality of our measures. For example, if we were going to make GAO personnel decisions on the basis of exit interviews, we would want to take steps to ensure that we were recording the real reasons why people leave, not just general and socially acceptable reasons. Hence, the initial validation plan should identify the concerns we had about the measures and initial plans for resolving these concerns. We cannot at this point complete the validation planning, because much of the validation and other quality assurance efforts require a completed questionnaire. However, the initial plan is important, because validation can be a major effort requiring extensive prior arrangements for field trips and the like. Without careful review and advanced planning, we may overlook a high-risk measure or not foresee the time and effort required to validate the measures.

Measure validation and data analysis are detailed and complex topics that require much discussion and the use of terms and concepts that we have not yet introduced. For these reasons, we have chosen to treat these topics briefly and in general terms in this section. More details and more complete explanations are presented in chapters 13 and 16.

In conclusion, the product of questionnaire planning is a documented description of all measures needed to implement the study. It is not a questionnaire. As we shall see in later chapters, to develop a questionnaire we must structure, organize, and translate these measures into a format appropriate for the respondent to read and answer. The questionnaire should include only the measures necessary to meet the project objectives. The measures should reflect the project needs for precision and generalizability, they should be structured in accordance with the logic needed to evaluate the required cross-sectional and temporal relationships, and they should relate to the units of analysis that will be used to draw conclusions from the data. In addition, the plan should include the documentation necessary to describe all standards needed for study comparisons. Initial analysis plans and plans for validating certain measures that are of concern should also be included.

CHAPTER 4

DEVELOPING THE SAMPLE FOR DATA COLLECTION

An important planning step in data collection is deciding what units to collect information from or about. There are two broad options: (1) selecting a sample so that it will be statistically representative of a larger universe of units and (2) selecting a sample without regard to the representativeness of the sample. We use the first option when we need to generalize from our sample to the universe, and we use the second when generalization is not a priority. In this chapter, we outline some of the main considerations in statistical sampling, in which the aim is to ensure that the sample is representative, and we briefly discuss some of the many nonstatistical methods that may be used on occasion.¹

STATISTICAL SAMPLING

The aim of statistical sampling is to use a sample from a universe in order to estimate the parameters of that universe. For example, if all the persons who participated in a certain government program constitute the universe, then the proportion of participants who are older than 25 is one of several universe parameters. Rather than determining that proportion by asking for the age of each person in the universe, we can draw a sample. We would determine the age of each person in the sample, compute the proportion of persons over 25, and then generalize from the sample to the universe. We now turn to the reasoning involved in generalizing from a sample to the universe.

To generalize our findings, we must first define the universe. In theory, we should enumerate every unit in the universe in a way such that every unit has an equal chance of being selected for the sample. In practice, it is unrealistic to expect to enumerate every unit in a real universe (such as services for the elderly or day-care centers), but the enumeration must be reasonably complete and accurate and be able to represent the actual universe. Second, we must draw a representative sample from this universe, and the sample must provide the degree of measurement precision and certainty generally accepted by the scientific community. Third, we must do this very efficiently. In many instances, accomplishing these three tasks will challenge the technical skills, creative abilities, and perseverance of the evaluation or audit team.

SURVEY UNIVERSE

We cannot determine the sample until we have studied the size and characteristics of the universe we want to know about.

¹For more detailed discussions of topics throughout this chapter, see Using Statistical Sampling, PEMD transfer paper 6.

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 universe is to learn about the universe distribution--the major categories of units and the numbers in each category. This step takes place even before listing the universe. For example, if we are sampling banks, we should learn the differences between county, regional, statewide, branch, and unit banking; we should know geographic location factors and understand the basis for classifying banks as very large, large, medium, and small. If we are studying unit commanders in the armed services, we should know the unit sizes and types and the variations among the services. This research will help us design sampling factors, such as stratification and stratification size, and will ensure a representative sample.

Once we are familiar with the universe's characteristics, we can look for sources that enumerate each unit in the universe. 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 impact that the Foreign Corrupt Practices Act had on U.S. business.² The universe was U.S. companies that conduct most of the foreign business. These companies were readily identified, because it is the "Fortune 1,000" companies that 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 universe.

In practice, we never have a list of the real universe; we have only a list at the time the source material was current. By the time we use our questionnaire, some units will have left the universe and others will have joined it. For example, in the "Fortune 1,000" evaluation, 6 percent of the firms left the universe, and we do not know who may have joined it. The sample analysis must evaluate and make statistical adjustments for the losses. Whenever possible, the impact of the additions should also be considered.

The best way to start enumerating a universe is to talk to experts in the field and search out likely organizations,

²The Foreign Corrupt Practices Act prohibits payments to foreign officials if the purpose is to influence business. The GAO evaluation of this act was reported as Impact of Foreign Corrupt Practices Act on U.S. Business, GAO/AFMD-81-34.

archives, directories, libraries, and management information systems until we discover a reliable source. Then we organize, reorganize, or index the sampling units or elements 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 users of military medical facilities. We had no problem locating active-duty users, but finding retired military personnel was a problem. Although each service had a reasonably accurate and current computerized list of the names and addresses of its retirees, we did not know which or how many were potential users at each hospital in the hospital universe. We also did not know how the information was coded on the tapes, and the tapes were incompatible with our address file system. We had to decipher the tape codes, translate them into a compatible format, and merge the three services' tapes. The next step was to find a way to associate individual names in the universe with individual military hospitals. Our field work showed that personnel 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. This effort took several trips to computer archives in various parts of the country and several weeks of computer programming.

In another study, a study of group homes for the mentally disabled, we discovered that group homes were, for all practical purposes, a hidden population. When we took our sample, we began by using a list of community mental health "catchment" areas stratified by urban and suburban areas. While these areas were smaller than standard metropolitan statistical areas (SMSAs), they were clustered around the SMSAs, and the clusters were analogous to SMSAs. The Health and Human Services directors of the catchment areas knew where the group homes were in their areas. We matched catchment areas to SMSAs and sampled locations. Using maps, we identified the municipalities, towns, and townships in the areas. After a few hundred or so phone calls, we eventually mapped the zones and identified the responsible officials for these areas. It took hundreds of phone calls to the catchment area directors to come up with an address list of 1,000 potential group homes. This list had to be culled because it included all potential group homes in the area, some of which did not fit our criteria. The end result, however, was a sample of geographic locations likely to have zoning problems. We also had the names and addresses of group home operators and zoning officials in each location. However, our sampling unit was locations, not operators or officials.

Sometimes, no matter how hard we search, we cannot find archival data or records from which to develop a universe. 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 we may want to look at existing data to surmise some ratio or relationship associated with the universe. For

example, if we want to define the universe of flight-service-station specialists and we estimate that the average number of specialists per station is 16, we can multiply 16 specialists by the 316 stations and estimate the universe at 5,056.

Similar methods can be used to estimate the number of parolees in a city. By sampling 35 parole officers and counting the number of cases each has, we can develop an average caseload. Say the average caseload number is 92. If we multiply 92 cases by the number of parole officers in the city, we have an estimate of active parolees. Sometimes ratios can help. For example, suppose our preliminary work indicates that one of every nine teenage girls in the nation is likely to have an out-of-wedlock pregnancy. Since there are 16 million teenage girls in the United States, the number of teenage out-of-wedlock pregnancies is likely to be 1.8 million.

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

In another 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 universe descriptions of their members. However, their cooperation varies with their interest in what we are doing. The cost for lists can be anything from a few hundred to several thousand dollars. Although the Census Bureau sometimes has lists we need, 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 Directory, Polk, and Thomas, sell universe lists. Also, some commercial firms 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 the buyer knows very little about how the lists 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. We did not test this list, and we got one of the poorest response rates (60 percent) in our 10-year history of survey work.

CONFIDENCE AND PRECISION

Once we have enumerated the universe and are sure that it represents the universe to which we want to generalize, we are ready to design the sample. Two important considerations are the precision and confidence levels. Precision, which in this case is an indicator of sampling error, is the maximum amount by which estimates of a universe parameter, based on the sample, are expected to differ from the true (real or actual) value of the universe parameter.³

Precision or sampling error may be stated at different confidence levels--90 or 95 percent, for example. The confidence level tells us how much we can trust our sample estimate of the universe parameter. We say, for example, that we are 95-percent certain that our estimate will not differ from the true value of the universe parameter by more than the sampling error. Or, to put it another way, the chances are 19 in 20 that our sample estimate will not differ from the true value by more than the sampling error.

Here is an example. Suppose we want to estimate the proportion of persons older than 25 who participate in a government program. Suppose, further, that the true value of the proportion in the universe is 50 percent. When we use a sample to estimate the proportion, we obtain a conclusion of the

³Besides sampling error, there are several other types of error, such as measurement error and data analysis error. All these sources of error combine to affect precision. However, a full treatment of this subject is beyond the scope of this chapter. Measurement error is discussed more completely in chapter 11 and in a forthcoming PEMD transfer paper on measurement. See also Using Statistical Sampling.

following form: the proportion of persons older than 25 participating in the program is estimated to be .51 (or whatever number is computed from our sample) plus or minus .05 (the confidence interval or sampling error) with a 95-percent confidence level. This means that we do not know exactly what the universe proportion is but (1) we estimate the number to be .51 and (2) if we were to take a large number of samples with the precision of the one just drawn, our confidence interval would cover the true value 95 percent of the time. The width of the confidence interval is what we mean by the precision of the sample or the sampling error.

Both the confidence level and the precision give us information about our sample estimate of the proportion of persons older than 25. However, the two indicators are interrelated. We can increase the confidence level simply by decreasing the precision. For example, we can report to the 99-percent level of confidence, but the sampling error will be plus-or-minus .10 rather than .05. Or we can increase the precision to plus-or-minus .02 by dropping the confidence level to 90 percent. However, GAO usually keeps the confidence level at 95 percent, regardless of the sampling precision.

SAMPLING ERROR

Sampling error is controlled mainly by sample size, but because sampling error is determined by five different factors that often interact, we need to consider the overall situation before focusing on sample size. In general, sampling error is determined by

1. the size of the sample,
2. the size of the universe,
3. the percentage of the population in the sample,
4. the stability and variability of our measure,
and
5. the proportion of the universe presenting
the trait we are interested in.

First, and simply put, the bigger the sample or the more cases we look at, the smaller the sampling error. Second, and conversely, the sampling error is likely to get bigger as the universe gets bigger. We say "likely," because there comes a point at which the universe is so big that it does not matter any more, and further increases will not change the sampling error. Third, the greater the proportion or percentage of the universe in our sample, the smaller the sampling error. Fourth, the more stable the measure, or closer the measures fall together, the smaller the sampling error. Conversely, the more dispersed or variable the measures are, the greater the sampling error.

So far, most of these concepts are intuitive, but the last consideration is more difficult. The sampling error is smallest when either a very large proportion or a very small proportion of the universe has the characteristic we are interested in. Suppose we are studying the proportion of the U.S. population that is younger than a specific age. The sampling error gets worse (larger) as we approach the median age or the age at which half the population is above and half below. This is why we assume a 50-50 split when we estimate sampling error; it is our worst-case condition. As we will show in the next section, these concepts are very important in determining sample size. We have presented a very brief and oversimplified discussion of some very important considerations.

CALCULATING THE SAMPLE SIZE

It is important that persons who want to determine a sample size thoroughly understand the complete project design. Regardless of what size is chosen, it will be bigger than necessary for some measures and analyses and smaller than necessary for others. And that is the point; the chosen sample size should be a logical compromise that accounts more or less completely for all the elements that must be covered in the audit or evaluation. Failure to make this compromise can either severely limit the audit or evaluation because of excessive undersampling or render it very inefficient because of excessive oversampling.

For example, in a study of employees who had been through a reduction in force, the initial request was to sample 300 people to see if they were still unemployed. However, it soon became obvious that the sample was too small. The overall project design also required information on the differences between cities, between city and county programs, and between blacks and whites, men and women, high-school graduates and non-high-school graduates, and black non-high-school graduates and the rest of the labor force. The final survey design called for a sample of 1,000. The precision varied from plus-or-minus .04 to plus-or-minus .15, and the confidence levels varied from 95 percent to 90 percent, depending on the use of the sample. In some measures and tests, the confidence or precision was greater than needed to prove our point, and in others it was marginal. Overall, the design provided the most cost-effective compromise for addressing all the information requirements laid out in the project design.

Cost effectiveness is a critical factor in determining sample size. As we stratify and cross-stratify to meet the project design requirements, the sample size explodes. For example, a matrix displaying 18 classifications by another 18 classifications, small for the typical GAO study, could require a sample size of 50,000. Hence, it is very important to ensure that all stratifications are essential to the evaluation's goals and will contribute to the report.

Another factor to consider is the sampling procedure. Usually, the first estimates of sample size are based on simple random sampling. However, the simple approach may not be cost effective. Fortunately, the discipline of sampling is sophisticated and includes techniques that are much more powerful, although they are more complicated. Examples are stratified samples, proportionate stratified samples, single and two-stage or multistage cluster designs, and ratio samples.

If the conditions are suitable, these techniques can provide the same level of confidence or precision with half the sample size of simple random sampling. The group-home zoning project illustrates the efficiency of these techniques. On this project, the design required the sampling of locations for group homes.⁴ By using a single-stage cluster design rather than a simple random sample, we were able to reduce the sample size from 300 to 100 and retain the same level of confidence and precision. This saved nearly \$100,000 and about a year of extra work.

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 we select our cases. We either study all the cases --that is, take a census--or select part of the universe in a nonstatistical manner. When we take part of the universe, we usually do so for a reason. It may be that we are doing a case study, so we select one or more cases that provide the best opportunity to observe the phenomena or relationships of interest, and we do not want to generalize our findings to the universe. In other situations, we know very little about the universe and we cannot draw a statistical sample, so we arbitrarily select as many cases as we can and report our findings. However, in many situations, we want to generalize, and we know something about the universe but it is just not feasible to draw statistical samples. So we pick a sample that we hope will correspond, in its features, to the universe, even though we know we will not be able to use the powerful reasoning associated with statistical samples. An important category of nonstatistical sampling is called "judgment sampling."⁵

A judgment sample draws its name from the fact that in the judgment of the authors, the cases chosen correspond to certain aspects of the universe. The cases may be selected because they

⁴The sampling design is documented in An Analysis of Zoning and Other Problems Affecting the Establishment of Group Homes for the Mentally Disabled, GAO/HRD-83-14.

⁵For a discussion of several forms of nonstatistical sampling, see W. E. Deming, Sampling Design in Business Research (New York: Wiley, 1960).

are judged most typical, because they represent the extreme ranges, because they represent a known part of the universe, or because they simulate or act as a proxy for a representative sample from the universe. 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 Texas, Mississippi, New York, and California, because these states represent the extremes of the laws and practices. We could study 50 prime contractors with the Defense Department in California and New York, because these contractors account for 82 percent of all DOD 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 will usually not be nearly as persuasive as with statistical samples. However, occasions may arise (as with a very homogeneous universe) in which the situation is not altogether bleak.

When the validity of our findings is dependent on the extent to which we can generalize them to the universe, and when we do not have a statistical sample, it might help if we had 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 universe. A very large, statistically valid random sample might give us a value of 9. A large, medium, and very small but statistically valid random sample might give us 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 universe 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 universe. (Incidentally, this is how we select our pretest candidates, because we do not have the time or 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, because these levels could be presumed similar for most

U.S. city-dwellers. However, in this review, we were concerned about programs and their effects, and the programs and effects 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 universe 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 universe of people in one of the 132 health maintenance organizations, the 16 case studies together permitted only very careful and limited generalizations. 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 5

FORMATTING THE QUESTIONS

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

OPEN-ENDED QUESTIONS

If data analysis did not have to be considered, open-ended questions would be used more often than they are, because they are easy to write and require very little knowledge of the subject. All we 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 incomplete and ambiguous answer, and it will be very difficult to use these answers in the 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 that others may or may not mention. Open-ended questions do not help respondents consider a range of factors; rather, they depend on the respondents' recall. Because we have 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 analyze the answers together.

To begin with, we 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. Are these answers the same, and can we group them? We have no way of knowing. "More convenient" or "less trouble" may refer to faster transit time, acceptance of shipping volume, regular availability of space, better schedules, more frequent sailing, quicker documentation, better support service, better sales service, or vessel-cargo compatibility. Such answers give us little ability to consider anything.

Another problem is that we cannot machine-process open-ended questions. We must use a complicated process called "content analysis," in which we read and reread a substantial number of the written responses, identify the major categories of themes, and develop rules for assigning responses to these categories. Then we have to go through the entire sample to categorize each answer. Because people categorize answers differently, three or four people have to categorize the answers, using rules to determine interrater reliability, before we get a working data base. Furthermore, we end up with no raw data to display and combine or aggregate, and the data base we get will have much less precision and much broader categories than that developed from closed-ended questions.

There are several other problems with open-ended questions. Because no structure is provided for the answers, people are uncertain about what the question means and how they are expected to answer. Hence, their answers vary greatly, as do answer lengths. Some respondents write a word, some a phrase or a sentence, and others a paragraph or a page. Still another problem is that open-ended questions substantially increase response burden. They 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. Heavier response burden, in turn, contributes to two other problems: the quality of the answers varies with the respondents' literacy, and people are 10 times less likely to answer open-ended questions. Our experience is that the nonresponse rate varies from 30 to 60 percent for open-ended questions.

Because we usually need very precise and factual evidence, we should use open-ended questions sparingly. But it does happen that open-ended questions are unavoidable when we are uncertain, for example, about criteria and must develop them and in other exploratory work. Also, open-ended questions do sometimes have advantages. If we ask enough people an open-ended question, we can develop a list of alternatives for closed-ended questions and will not have to do as much research. We can also use open-ended questions to check the quality of our structured alternatives, to make sure we do not miss items, and to give the respondent a chance to mention items that may be less relevant to the population but are very important to the respondent. Open-ended follow-up questions are frequently used to ensure complete coverage. For example, we may offer a yes/no question and follow it up with "Why 'no'?" At the end of a questionnaire, we use an open-ended question to give respondents a chance to comment on any of the items in the questionnaire.

The rest of this chapter deals with closed-ended questions, because they are the meat and potatoes of our work. They have to be just right. That is, all relevant alternatives must be listed and each alternative must be necessary, plausible, and analyzable. Leaving out relevant alternatives may cause over-reporting for some choices and underreporting for the omitted choices. Research and testing are needed to develop a sound list of alternatives.

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

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 here and on page 50.

Row format

What is the estimated number of children, juveniles, or adults that you usually care for at any one time? (Answer for each appropriate age group.)

| <u>Age group</u> | <u>Number of children, juveniles, or adults</u> |
|--------------------------------|---|
| 1. Under 6 years of age | _____ |
| 2. From 6 to under 9 | _____ |
| 3. From 9 to under 12 | _____ |
| 4. From 12 to under 14 | _____ |
| 5. From 14 to under 16 | _____ |
| 6. From 16 to under 18 | _____ |
| 7. From 18 to under 21 | _____ |
| 8. Adults over 21 years of age | _____ |

Column format

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, number of exploratory wells, and number of developmental wells.

| A. Country | B. Exploration activities | | | | |
|---|--------------------------------|-------------------------------------|--|-----------------------------|-------------------------------|
| Countries in which you are doing exploratory work | Square miles under exploration | Line miles of seismic data gathered | Square miles of area under drilling rights | Number of exploratory wells | Number of developmental wells |
| | | | | | |

Matrix format

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

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

| | City | County | State | Other |
|-------------------|------|--------|-------|-------|
| 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 also very good for filters in the line of questioning and can be used to move respondents to the questions that apply to them. For example we might ask

Did you get training?

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

But most of the questions we ask deal with measures that are not absolute or measures that span a range of values and conditions. An example is "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. Third, parts or some contracts may have been clear and others not clear. (For instance, one contract may have been clear with respect to the products to be delivered, the packaging and shipping requirements, the delivery date, the cost, and the insurance but unclear with respect to the test, performance, and

workmanship standards as well as the contract's effective date.) In short, we need at least 12 yes/no items to answer this simple question.

Furthermore, yes/no questions often do not give much information. For instance, a yes answer to "Was there community opposition or not?" tells us little. We have no idea how much opposition was involved; it could have been insignificant or overwhelming.

Yes/no questions are difficult to write because we must be very precise to discourage quibblers. Take the following situation. We ask respondents, "Do you think this regulation is clearly written?" If text of the regulation is generally clear but some phrases are a little awkward or ambiguous, some people may answer "no" because some phrases are unclear; they might not consider the text as a whole.

Because we have to be very precise and because we get so little information from yes/no questions, they are very inefficient. We must administer several rounds of questions 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 leads to serial repetition and 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 or got site support but not mission training?

Other question-writing mistakes are also common--for instance,

Did you get mission and 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

This example has several problems. The question and the response list do not agree; one is a yes/no format and the other is a multiple-choice format. 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 construction creates an unintended double negative. But deleting the "no" to

correct this also does away with the yes/no format. Such questions confuse respondents, add to their burden, and may 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 you do this, you must qualify or avoid the use of yes/no choices in the answer. If you do not take 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." Hence, the "no" answer could leave some readers worrying over a double negative. To avoid confusion, either qualify the answer choices or prevent the use of "yes" and "no" in the answers. The question and answer could read as follows:

Do you submit reports or not?

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

IMPLIED-NO CHOICES

In the following question, failure to check an item implies "no." We use the implied-no choice format because it is easy to read and quick to answer.

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 we want to emphasize the "no" alternative, we expand the implied-no format to include one column for "yes" answers and one for "no." We list "no" as an option when the respondent might not answer or might overlook part of the question. These omissions occur when the choices are difficult, the list of items is long, or the respondent's recollection is taxed. If we do not include "no" as an alternative, we will overreport no's, because we will not be able to differentiate real no's from omissions and nonresponses. An example follows:

Did the VA ask if you had the following health problems during or since your service in Vietnam? (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 | | |
| 13. Other (describe) | | |

SINGLE-ITEM CHOICES

In single-item choices, respondents choose not "yes" or "no" but one of two or more alternatives:

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

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.

On the positive side, the differences between the choices are usually clear, and we 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 skip them through parts of the questionnaire. Because the single-item choice does not have the popular appeal of the yes/no question, it is not likely to be overused and cause excessive serial repetition. Furthermore, the question writer is not likely to compress the question into a double-barreled item or to offer the respondent multiple-choice answers. The single-choice format is also not subject to bias from yea sayers or naysayers. And eliminating the negative alternative reduces misinterpretation.

But there are other 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 the tendency, may overcompensate and wind up overemphasizing the second alternative. This presents difficult problems for the analysis.

ENFOLDED FORMATS

One way around the yes/no constraints is to use an enfolded format:

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

The enfolded format gives us a measure of intensity, avoids some of the biases common to yes/no and 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?" Many students who could not have made it through college without a loan would be reluctant to admit this. However, they would say "probably yes" and perhaps even "probably no," given these options. Also, some of the borderline cases would check either "probably yes" or "probably no."

Enfolded scales do not usually change the overall proportions of affirmative and negative choices, but they give better measures of intensity. Furthermore, because people have less trouble making a choice, they answer enfolded items much more quickly and with fewer mistakes than they do the standard 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 and single-choice, and enfolded 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 give us error, bias, and unreliable answers. To avoid this problem and to reduce the respondent's burden, we can add a middle category:

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

The proportion of yes's to no's will not change, and we will have a better measure of the yes/no polarization, because the middle category absorbs those who are uncertain. However, the middle category may introduce a bias that favors either the first or last option. These problems can be avoided by careful item design. A good rule of thumb is that if we are not certain that nearly everyone can make a choice--that nearly everyone is in either one category or the other--we include a middle category.

Usually, the question asker will put in an "escape clause" 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." The format might look like this:

1. ☐ Yes
2. ☐ Probably yes
3. ☐ Uncertain
4. ☐ Probably no
5. ☐ No
6. ☐ Have not considered the issue

MULTIPLE-CHOICE FORMAT

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

What reasons best explain why you or your family went or had to go elsewhere for care? (Check one or more.)

1. ☐ No doctor(s) was(were) available to treat your particular case.
2. ☐ There was a very long waiting list for an appointment, so you were advised that it was better to go elsewhere.
3. ☐ The equipment required for your care was not available at that facility.
4. ☐ The facility was very busy and you preferred to go elsewhere for care.
5. ☐ Other (describe) _____

Multiple-choice questions are difficult to design because the writer must provide a comprehensive range of nonoverlapping choices. They must be a logical and reasonable grouping of the types of experiences the respondents are likely to have encountered. Hence, the writer must know the respondent population. There should be no doubt in the respondents' minds about how they should answer.

The real example above turned out to be flawed. We learned during the pretest that we had left out some important choices, such as "other facilities provided better care" and "needed continued care with the same physician." As a result, too many respondents had to write in why they went elsewhere for medical treatment.

Because this format is very important and requires the most research, field work, and testing, we discuss multiple-choice question design in chapter 8 in considerably more detail.

RANKING QUESTIONS

As the name implies, "ranking formats" are used to rank options with respect to their priority, importance, size, or cost. That is, we ask respondents to tell us which alternative is the highest priority, which is the second highest, and so on. They rank the choices with respect to each other, but their answers tell us little about the intrinsic value of their choices.

Ranking formats are difficult to write and difficult to answer. They give very little real information and are very prone to errors that can invalidate all the responses. They should be avoided whenever possible in favor of more powerful formats and formats less prone to error, such as rating.

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. Lawn care, telephones, and choir practice would probably be a distant third, fourth, and fifth.

Ranking is hard for people when there are more than five categories. This is because for five items, they can easily pick the first and second and then the last and next to the last, so that what is left is the middle. But for more than five 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.

Ranking-questions have to be written very carefully. The slightest lapse in clarity will cause people to rank in the reverse order or rate rather than rank. Without clear instructions and a well-designed layout, respondents will assign two alternatives the same rank or forget to rank every alternative.

Sometimes ranking must be used. The example presented on here and on the next page is one that has worked reasonably well. Respondents will make a few errors, but we have statistical procedures to handle them.

The demonstration of program results is obviously an important factor in the continuance of program funding from the Office of Education (OE). 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 OE or state education agency program (SEA) officers? Indicate your answer by rank ordering each of the following alternatives from the most to least impressive. Select the type of results which you think is most likely to impress OE or SEA officials. Rank this 1st by circling. Do the same for all the remaining categories, ranking them 2nd, 3rd, 4th, 5th, 6th, and 7th.

- | | | | | | | | |
|--|-----|-----|-----|-----|-----|-----|-----|
| 1. Improvement in educational management or accountability | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th |
|--|-----|-----|-----|-----|-----|-----|-----|

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

Although this ranking format can structure several alternatives, it becomes more complicated if we try to go beyond this. If we have, say, 12 to 15 factors, it is better to take another approach. We ask respondents to write down the five most important factors, then the five least important. They can usually handle the top and bottom alternatives but have trouble in the middle. (The middle will be those that are left over.) Next, we ask the respondents to go to the first grouping and rank the first and second most important alternatives 1 and 2 and the least and next to least important 5 and 4, respectively. If the respondents repeat this procedure for the middle and end groups, they will provide all the information the evaluator needs to rank all categories.

Ranking should be used only for alternatives that are so close in value that most people would classify them in the same value category. And this is very rare. In most cases, including the example here, we can still make the necessary distinctions and get a better level of quantification with rating.

RATING QUESTIONS

As explained earlier, ranking values are assigned solely on the basis of relative order--first, second, and third. This applies regardless of the measures' absolute position within a range of possible values. Ratings, however, are assigned solely

on the basis of the score's absolute position within a range of possible values--of little importance, somewhat important, moderately important, and so on.

Here are two examples of rating formats:

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

Under what risk classification should Presentence Investigation reports contain recommendations for court conditions?

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

As can be seen, rating scales are easy to write, easy to answer, and provide a level of quantification that is adequate for most purposes. If they are used in appropriate circumstances, they produce reasonably valid measures.

However, on occasion, we may be concerned about the fineness of discrimination. Suppose, for example, we feel that most people will give high ratings to all categories. How will this help us distinguish the importance of the various categories? A survey on delinquent and neglected children illustrates the point. If we asked respondents to rate the importance of various services--health and development, mental health, education, vocational education, family and diagnostic services, and drug and alcohol use--they would tell us that all are essential. The problem can be overcome by asking a rating question as a filter and a ranking question for the fine discriminations between the high-priority items, as in the examples on the next page.

What are the priority needs of the children or juveniles in your institution? Do not consider whether or not your institution has the capability or authority to address these needs. (Indicate your answers by checking one and only one priority rating column for each priority need.)

| Priority needs | Of little or no importance | Somewhat important | Moderately important | Very important | Essential |
|--|----------------------------|--------------------|----------------------|----------------|-----------|
| 1. Health and development services | | | | | |
| 2. Mental health services: social, psychological, psychiatric, and counseling services | | | | | |
| 3. Educational (academic) services | | | | | |
| 4. Vocational services | | | | | |
| 5. Family services | | | | | |
| 6. Diagnostic services | | | | | |

Consider only the needs that you checked as essential. Now rank the essential needs by order of decreasing importance. Do this by selecting the most important of all the needs you considered essential. Rank this 1st by circling. Do the same for the remaining essential needs, ranking them 2nd, 3rd, etc., until you have ranked each of the needs checked as essential. (REMEMBER EACH NEED CAN HAVE ONLY ONE UNIQUE RANKING SCORE. Check for this by making sure you do not have more than one circle in each row or column.)

| Priority needs | Ranking |
|--|-----------------------------|
| 1. Health and development services | 1st 2nd 3rd 4th 5th 6th 7th |
| 2. Mental health services: social, psychological, psychiatric, and counseling services | 1st 2nd 3rd 4th 5th 6th 7th |
| 3. Educational (academic) services | 1st 2nd 3rd 4th 5th 6th 7th |
| 4. Vocational services | 1st 2nd 3rd 4th 5th 6th 7th |
| 5. Family services | 1st 2nd 3rd 4th 5th 6th 7th |
| 6. Diagnostic services | 1st 2nd 3rd 4th 5th 6th 7th |
| 7. Drug/alcohol abuse services | 1st 2nd 3rd 4th 5th 6th 7th |

GUTTMAN FORMAT

In the Guttman format, the alternatives increase in comprehensiveness; that is, the higher-valued alternatives include the lower-valued alternatives. Consider the arithmetic operations of addition, multiplication, and division. We generally assume that a person who can multiply can also add. If you can divide, you get credit for all three operations.

In one job, we asked state resource officials how they benefited from an earth-orbiting satellite. The question was

Identify the benefit areas and the degree to which you can qualify and/or quantify these benefits. (Check one column in each row.)

| Benefit area | No benefit identified | Identified benefit | Measured benefit | Measured secondary and disbenefits | Assessed primary and disbenefits | Value of dollar made of benefits analysis |
|--|-----------------------|--------------------|------------------|------------------------------------|----------------------------------|---|
| 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 (specify) | | | | | | |

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 disbenefits, as well as the worth or dollar value of these benefits and disbenefits.

Guttman formats can be very useful when we have to summarize a very pluralistic assessment. For example, in the illustration, instead of having six independent indexes, we could have had one index that subsumed all the others.

LIKERT AND OTHER INTENSITY SCALE FORMATS

Likert and other intensity scale formats are usually used to measure the strength of an attitude or an opinion. Here is an example of the Likert format:

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

Other formats use "oppose" and "support" or "pro" and "con."

Intensity scale formats are very easy to construct. All the question writer has to do is make a statement and follow it with an agree-or-disagree response choice. But Likert formats are also subject to bias, because the statement presents only one side of an argument. We may know that respondents agree with the statement, but we must infer what they believe. We can learn more from either of the following formats:

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

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 and similar scales are best used when respondents can agree or disagree with an actual policy:

Some people agree with GAO's policy on rotation, while others do not. The question is, How do you feel about the policy?

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

Intensity-scale items are subject to the natural tendency of people to agree. Likert users counter this by presenting the converse statement also. For example, they would ask for a response to "My boss does not let me participate in decisions (agree or disagree)." Then they would ask for a response to "My boss lets me participate in decisions (agree or disagree)." However, they would have made two biased statements and still not have a measure of the respondents' perceived participation in decisionmaking. It would be better to ask, "To what extent, if at all, do you participate in management decisions?" Now we have a measure of what respondents think they do. The point is to use a format that allows us to directly quantify the strength of the respondents' actual attitudes or opinions rather than using an indirect approach that quantifies the respondents' reaction (agree or disagree) to a statement that in turn is supposed to quantify a measure (the measure here is participation).

Quantifying amounts and frequencies

Many questions ask the respondent to quantify either amounts or frequencies. These formats are relatively simple. They use adjectives and adverbs to describe the amount, frequency, or number of items they are measuring. Some amounts can be described as help, hindrance, impact, increase, or decrease. Others can be quantified as little, some, moderate, great, or very great. Sometimes such adverbs as "very" and "extremely" are used. We can also imply quantities by the sequencing of our answers:

1. ☐ Little or no hindrance
2. ☐ Some hindrance
3. ☐ Moderate hindrance
4. ☐ Great hindrance
5. ☐ Very great hindrance

When we list five options, answer 3 is the middle of the scale, answer 1 is one extreme, and answer 5 is the other extreme. It is logical to assume that answer 2 is a quarter of the way on the scale and that answer 4 is three quarters of the way. Frequencies or occurrences of events are treated the same way:

1. ☐ Seldom if ever
2. ☐ Sometimes
3. ☐ Often
4. ☐ Very often
5. ☐ Always or almost always

We might also anchor frequencies with fractions or percentages:

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)

Or we might use verbal descriptions:

1. ☐ Very often (once every other day)
2. ☐ Not very often (once a month)

Or descriptions and words that divide the range into intervals:

To what extent, if at all, does your water pollution monitoring system cover the streams in your county?
(Check one.)

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, all or about all of the streams are covered

Intensity-scale guidelines

Some guidelines for using intensity scales follow.

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 only the law can 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-numbered scale, preferably with five categories.
3. If there is a possibility of bias in the scale order, order the scale in a way that favors the hypothesis you want to disconfirm and disadvantages the hypothesis you want to confirm. This way, you confirm the hypothesis with the bias against you.
4. If there is no bias, put the most undesirable choices first.
5. Pick a scale range and anchor (that is, specify the ends of the range) with concrete and unambiguous measures.
6. Use the item sequence and numbering to help define the range of categories.
7. Use words that will divide the scale range at appropriate intervals.
8. Anchor the intervals with numbers, fractions, or proportions and descriptions, when feasible.

SEMANTIC DIFFERENTIAL FORMAT

In the semantic differential format, the values that span the range of possible choices are not completely identified; only the extreme values are labeled. An example is

Indicate the number of times per week you initiated technical communications with colleagues in your group.

Few

Many

(2 or 3) ☐ ☐ ☐ ☐ ☐ ☐ ☐ (20 or more)

The respondent must infer that the range is divided into equal intervals. The range seems to work better with seven categories rather than the usual five. The reasons for this are complicated, but seven categories provide a closer approximation to the normal distribution.

Semantic differentials are very useful when we do not have enough information to anchor the intervals between the poles. However, three major problems detract from this format. First, the questions are very difficult to write well, and if they are not written well, respondents will not answer or will answer with errors. Second, because the semantic differential has no midrange or intermediate anchors, respondents will flounder and are more apt to make judgment errors. Third, the results lack a certain amount of credibility because they are not tied to a factual observation. For example, there is a big difference between saying that 70 percent 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.

PAIRED COMPARISONS AND CONSTANT REFERENT COMPARISONS

Two other formats are occasionally used: the paired comparison and the constant referent comparison. In the paired comparison format, the respondent is asked to do a series of comparisons:

Consider the following four rotation policies: (1) internal rotation, (2) intergovernment rotation, (3) external rotation, and (4) no rotation. Do you prefer internal rotation or intergovernment rotation? (Check one.)

1. ☐ Internal rotation

2. ☐ Intergovernment rotation

Do you prefer internal rotation or external rotation? (Check one.)

1. ☐ Internal rotation
2. ☐ External rotation

The same type of question is asked until all comparisons have been made. For four alternatives, we need six comparisons. The number of comparisons increases exponentially with the number of choices.

Paired comparisons are cumbersome, and the mathematics necessary to develop a scale is difficult. Therefore, this format is used infrequently. However, paired comparisons do provide a very powerful measure.

Constant referent comparisons use one standard and compare every choice with this one standard. An example is in a comparison of ranger training, counterinsurgency training, and jump training with basic training:

Which training was more difficult? (Check one in each row.)

- | | |
|--|--|
| 1. <input type="checkbox"/> Basic training | <input type="checkbox"/> Ranger training |
|--|--|
-

- | | |
|--|---|
| 2. <input type="checkbox"/> Basic training | <input type="checkbox"/> Counterinsurgency training |
|--|---|
-

- | | |
|--|--|
| 3. <input type="checkbox"/> Basic training | <input type="checkbox"/> Jump training |
|--|--|

CHAPTER 6

AVOIDING INAPPROPRIATE QUESTIONS:

THE IMPORTANCE OF PRETESTING

To make sure our questions are appropriate, we 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 while the converse was true for the obviously handicapped.

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

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

The best way to avoid inappropriate questions is to know the respondent group and not rely on stereotypes. A brief story 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 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, one survey asked military personnel for their MOS

numbers (numbers identifying their military occupational specialties). Because the survey was to assess whether people worked in their occupational specialties, this question was vital. However, during pretesting, we found that many Marines did not know their MOS numbers. Either they did not answer (in which case we did not know which specialty to relate their comments to) or they gave us the wrong number (in which case we attributed their comments to the wrong occupational specialty). In another survey, we asked for information for 4 years, but the respondents kept records for only 3 years. In still another, we asked an agency for cost accounting and indirect labor information it did not keep.

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 performance of their units, whereas enlisted personnel were more likely to "tell it like it is." Similarly, physicians in European military hospitals rated the quality of their own medical practice very high but were critical of their 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 got much more accurate observations from other sources (enlisted members of the units and peer and nurse reports.)

Sometimes respondents provide misinformation because they make a random guess. Some people do not like to admit that they do not know something. Some like to please the question asker by responding. But we want true observations, not guesses. The problem is that sometimes we do not know who in a population is likely to have the specialized knowledge or experience and who is likely to guess. For example, we may not know how to locate people who could not get the care they requested from military hospitals.

There are two approaches to locating subpopulations with special characteristics. We can develop procedures that direct the questionnaire to the most knowledgeable people in the population, or we can get people who do not have firsthand experience to select themselves out.

In the following examples, we developed procedures to direct questionnaires to knowledgeable respondents. One project evaluated the usefulness of a publication that analyzed federal funding by type of program and geographic location. The congressional staff members we expected to find using this report did not in fact use it regularly, but other congressional staff did. We analyzed staffing patterns and developed procedures to direct the questionnaire to the right staff members. In the other project, we wanted to survey supervisors of handicapped employees. We used the job description of each handicapped employee and organization charts to identify this population.

We use self-selection techniques when we cannot find a practical way to direct questionnaires to the right people. For example, in a child-care survey, we asked parents to respond to a small part of the survey if they did not have children of day-care age, in this way selecting themselves out from the rest of the survey and future inquiries. In a survey of the disabled population, we asked only the people who felt they could benefit from special accommodations to answer certain questions. In a survey of military hospital clients, we sent a short mailgram questionnaire asking if they had been unable to get the service they needed. Those who answered "yes" were sent a more detailed questionnaire. We can also get the right people to answer the right question by asking respondents to skip certain questions.

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

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

Another means of selection is to ask people to rate their expertise. For example, we asked state policymakers to rate their knowledge of policies that affected certain special populations. In a study of the feasibility of a national health plan, we asked people to rate their expertise in the health care industry, insurance, education, manufacturing, preventive medicine, and so on.

QUESTIONS THAT ARE NOT GEARED TO RESPONDENTS'
DEPTH AND RANGE OF INFORMATION
AND PERCEPTIONS

One way to avoid this type of question is to not use words or terms the respondents 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."

We 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. To computer workers, a clean room meant an office that was clean, but to the staff who mounted computer chips, it meant a room that

was a "million clean" (a very high level of cleanliness based on the number of dust particles per square inch). 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.

We must be familiar with our population, and we 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 and the computer compatibility of tapes from the LANDSAT earth-orbiting satellite. (The tapes provide data used to make computer maps of 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 demonstration programs. We found that the officials were aware not only of the federal programs but also of many state programs. With this in mind, we expanded the line of questioning to include many of the important state demonstration programs. However, when we did this, we overstepped the respondents' capabilities; they could identify programs from their own states but not other states.

As the preceding example demonstrates, it is just as easy to assume too much as it is to assume too little. We 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. The distinction between federal, state, and local personnel was important to us, but respondents saw them all as "government men." In asking training officers about the evaluation techniques they used, we overdesigned the question by including choices on test-retest, single-blind, and double-blind experiments. The officers did not even understand our terms, let alone use them. In another evaluation, we asked mathematics and science teachers to add up a few numbers and calculate some percentages. We did not pay much attention to the instructions, and we did not pretest the item, because we assumed this population would have little trouble with simple arithmetic. This was a big mistake.

If arithmetic questions perturb math teachers, imagine what they do to others. Questions like the following, which require respondents to take a percentage of a percentage, are particularly difficult:

1. By approximately what percentage, if any, has your average fee per client increased since the Tax Reform Act of 1976?

_____ (% increase in average fee per client)

2. Of this increase per client, approximately what percentage would you attribute to the 1976 act's provisions regarding paid preparers?

_____ (% increase in average fee per client due to act's provisions)

It is also important to make sure that respondents' perceptions match ours. If we ask people from rural areas about a very large company, they 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)." To illustrate the size of a trillion dollars, we might add "the value of one third of all the economic activity in the United States for 1 year."

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 likely not 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.

When a question seems irrelevant but is not, we should explain why it has been included. In reviewing the effectiveness of area agencies on aging, we asked a series of questions on nutrition programs. We were very careful to explain why these questions were needed.

Frequently, however, someone asks us to include what is called a "rider"--an unrelated question for use in another evaluation. If we include riders, we have two problems. The evaluation now has a dual purpose. But how do we explain the dual purpose to readers? Second, we have to weave the questions into the questionnaire so that they do not seem irrelevant.

Aside from riders, there are three ways in which irrelevant questions 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 evaluation to cover their bets. They were not really interested in precise information but, rather, conducted the evaluation to be sure they did not miss anything.

Not one of these reasons is acceptable. The use of evaluations for such purposes wastes both time and money. In the first case, we seem to others to be on a "fishing expedition." In the second, we probably will get a great deal of the sort of information we do not need and not enough of what we want. The third case rarely leads to the development of useful information.

QUESTIONS THE 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. As discussed in chapter 12, questions should proceed in a logical order set up by the instructions and clearly denoted by headings and lead questions. 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 would 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 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 DOD; 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. We also had to provide a few more size categories than we actually needed, because many respondents expected company size to be broken out this way.

QUESTIONS THAT REQUIRE UNREASONABLE EFFORT TO ANSWER

We should avoid asking questions that require unreasonable effort to answer--that is, unreasonable amounts of time or work, 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 we are willing to do to help them, and what benefits they will get from participation.

In general, we try to keep form completion time to under a half hour. We can exceed this 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; respondents see that GAO has done all it can to keep the burden down and is willing to help; and respondents see the need for and value of the information.

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, we must be candid about it and tell respondents how long it is likely to take. If the form is properly designed, it can be completed much faster than respondents imagine. Pretesting is the only sure way to find out the completion time, the task burden, the difficulty of the questionnaire, and the respondents' willingness to accept the burden. The price is very high for a miscalculation. If we underestimate the burden, we may increase the nonresponse rate, get inadequate answers, and lose our credibility; if we overestimate, we may unnecessarily compromise the design to gain the acceptance of its users.

For example, one request for participation was answered by many letters of refusal, because we asked respondents to do many laborious calculations in reworking major parts of their tax returns. Unreasonable effort was also required for a questionnaire that asked each respondent for 34 accounting and audit policies identified and indexed by page number.

Complicated response formats can also be very burdensome. We 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 also should be avoided.

EMBARRASSING QUESTIONS

Questions that are embarrassing, threatening, personal, or sensitive should be avoided. We should not ask respondents to

disclose legal actions, behavior that makes them look less than ideal, or medical, financial, or personal problems. If we must ask questions of this nature, we should do so in a way that makes respondents at least minimally comfortable.

For example, in a child-care needs assessment survey, we asked for information on marital status. This question was sensitive because some of the mothers 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
2. ☐ Separated/divorced/widowed/never married

Approaches for dealing with sensitive questions are presented in more detail in chapter 11.

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. Because unclear writing is covered in chapter 7 and overlapping response choices in chapter 8, this section will focus on the other causes of ambiguity.

Improper qualification

Improperly qualified requests do not adequately specify the conditions or the observations we want respondents to report on. If we ask a report user if the report was timely, the user does not know if we are asking whether getting it took too long or it arrived after it was needed or both. We must specify. Improperly qualified items are a major frustration to question answerers and question askers alike. Question answerers are frustrated because they do not know how to answer, and question askers are frustrated because they get either no answers or answers they may not be able to use. Correcting these types of flaws is a major part of the questionnaire design and test effort.

The guidelines for correcting these shortcomings are broad, but they are nonetheless a good place to start. First, we should get to know how the respondent population talks, thinks, and does things. Second, we should try to make sure all our terms are as well qualified as we can make them. Third, we should look for problems when dealing with processes, sequences, sources, times, goods and services, organizations, classifications, functions,

disciplines, regions, programs, systems, space, business, government, and infrastructures. Fourth, we should substitute concrete terms or examples for abstract concepts. Fifth, we should make as few assumptions as possible, even for the most obvious terms, and then we should test to ensure our assumptions are valid.

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, and holding companies, this assumption was false and many managers could not answer. They needed to know what we meant by an "establishment." After a few weeks of testing, we finally came up with a solution that worked:

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

Several additional examples are presented here to further illustrate why some very basic terms need qualifying. In a 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. 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 the staffs lacked

audit experience. This problem cleared up when we realized that some of the inspectors general did not consider investigations and inspections as audits. The measure of audit experience increased considerably when we changed the question to read, "How many years of experience have you had with the government doing audits, investigations, or inspections?"

Abstract concepts

Abstract concepts, like poorly 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." These questions are difficult, because respondents cannot readily describe or quantify their observations of love, goodness, autonomy, or stability. We must help them.

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

1. present the concept as behavior,
2. provide more-concrete definitions,
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. The question read

"Indicate the extent to which you felt bothered by the following uncertainties:

1. Not knowing what the upper classes expected of me.
2. Not knowing what the officers expected of me."

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. We must 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 must 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.

Rating neighborhood stability was another seemingly simple concept that required extensive analysis. 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 blue collar residents, number of white collar residents, blighted housing, 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 questions expose respondents to high 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 we must ask these types of questions of these people because they are the best or only source of information. When this occurs, we should be careful to safeguard the respondents' identities and try to prevent any administrative or other uses that would have repercussions on our informants.

For example, we found that many group homes for the mentally disabled would be placed at risk if the information they provided were crossreferenced with the information provided by local zoning officials. Therefore, we corroborated their reports by using other methods that did not put them at risk. In another case, we asked employees who collected health insurance for physical examinations to cooperate in evaluating the claim audit procedure. The data showed that carriers were not thoroughly auditing claims. We placed the employees at risk because the carriers wanted the names of these employees so they could retaliate rather than correct their procedures. Of course, we prevented the disclosure of names.

We should not ask for proprietary information unless it is essential to the evaluation. By "proprietary," we mean information on new products, advanced designs, marketing strategies, compliance hearings, equal employment opportunity cases, financial data, national security information, and other restricted information. If we need this information, we should initiate safeguards and maintain a high resolve not to disclose it. GAO can be very proud of its record on this issue.

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, we can ask questions to get background, achievement, and behavioral information without asking respondents to take a test.

We should also avoid using questionnaires for administrative or entrapment purposes--that is, getting respondents to disclose self-incriminating information and then using this information to punish them. This is a very unfair practice, particularly when used on a population that is not wary (for example, the elderly or the mentally handicapped). Such tactics are counterproductive and compromise our data-gathering efforts. If we must gather this type of information, we should be entirely candid about it; that is, we generally tell the respondents that the information may be used against them.

UNBALANCED LINE OF INQUIRY

We should not write questions that develop a one-sided line of inquiry to support a particular position or preconceived idea at the expense of evidence to the contrary. The purpose of questionnaires is to develop information for an evaluation, not to make a case or win an argument. To do otherwise, in the long run, threatens the evaluators' reputation for objectivity, commitment to balance, and integrity.

CHAPTER 7

WRITING CLEAR QUESTIONS

To help potential respondents understand our questionnaires, we must write clearly and at the respondents' 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.

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, increases the reading level by about one grade. After this, every word increases the reading level by one year. Hence, 25-word sentences tend to make for effortful reading.

SIMPLIFY THE WORD STRUCTURE

Aside from shortening the sentence length, the most effective way to increase readability is to simplify the word structure. Four word structure factors affect readability: the size 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 six 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, the sentence is hard to read. For very easy reading (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 difficult to read if the ratio of root words to words with prefixes and suffixes is only 2 to 1. Reading becomes easier as we increase this ratio. Having four times as many roots as prefixes and suffixes makes for easy reading.

Finally, words used less frequently 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, readability should be tested. Several readability indexes focus on sentence length, word length, number of syllables, and word prefixes and suffixes. 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 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.

When we try to improve the readability of questions by using more-familiar words, we often use words with multiple meanings. 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 preferred 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, knowing there are no 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, phrasing concrete questions is very difficult. Furthermore, an undue emphasis on concrete words may cause an overly detailed, inefficient line of questioning. Therefore, it is important to choose the appropriate level of abstraction. It is a rule of thumb that the lower the level of 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. Less-literate people who have to concentrate on understanding the language may not be able to grasp complex ideas presented in this fashion. For instance, consider the following question:

"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 the question above might read as follows:

- "1. We need to know what proportion of your training followed a published schedule.
2. First consider the mission training you received during the unit assembly. What percentage of this training followed a published schedule?
3. Next consider the mission training received during annual active duty. What percentage of this training followed a published schedule?
4. Now forget mission training and concentrate on military occupational speciality (MOS) training. Consider the MOS training received during the unit assembly. What percentage of the occupational specialty training followed a published schedule?
5. Finally, consider the MOS training received during annual active duty. What percentage of this training followed a published schedule?"

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 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.

Sentences can be simple, compound, complex, or compound-complex. The simple sentence, containing a clear subject-verb relationship, should be our goal. However, because of the need for modifiers, qualifiers, and variety, the 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 of the sentence. If this is not possible, it should be at the end. Embedding the main idea in the middle of the sentence should be avoided, and 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 in the passive. In active 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 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, we 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

Because readers learn and rely on the structure used to develop sentences, 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 we are presenting. 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:

"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."

"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 can become monotonous and irritating.

Because people have trouble with even the simplest 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.

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. When we misplace modifiers, for example, we confuse the reader. Question writers also rely heavily on pronouns for concise writing. However, pronouns are often placed where they can modify 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. We get a similar problem when we use the word "which" 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?" A "yes" answer to this question did not give us the information we needed. We wanted to know whether people who currently did not have children younger than 12 living with them expected to have children younger than 12 living with them 2 years later. But because the antecedent of "this" was unclear, some people felt that a yes answer meant that they did not have children younger than 12 living with them and did not expect to have them in the future.

CHAPTER 8

DEVELOPING UNSCALED RESPONSE LISTS

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, such as for degree of importance or satisfaction.

Although this response format appears easy to write, it can be difficult and time-consuming. 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 specificity of the categories must be tailored to the measurement area.
- Respondents must feel that the order in which the categories are presented is logical.
- The lists should usually not exceed five to nine categories, unless the categories are grouped into sets.
- A 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, if we omit an important alternative, we will not get an accurate count for it.

To devise an exhaustive list, we must thoroughly research and understand the particular process or condition we are asking about. In many cases, pretesting is invaluable for ensuring the adequacy of the response list, because our respondent population usually knows the area better than we do. For example, in asking former DOD employees what DOD sources provided them with

information about postemployment restrictions, we forgot to include publications on retirees. Pretesting showed this was an important category.

The quality of medical care, organizational efficiency, and similar broad topics can be measured in a variety of ways. Developing comprehensive lists to measure such topics can be difficult. For example, the following question was used in evaluating veterans' satisfaction with agent orange examinations provided by VA medical centers:

Did the VA give you the following laboratory tests as part of your agent orange examination?

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, we avoid using the separate categories "single" and "divorced or separated." The word "single" can apply to either divorced or separated people. Another example of overlapping categories is

What is your occupation? (Check one.)

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

Because the categories are not sufficiently qualified, they are not mutually exclusive. Managers, technicians, secretaries, and sales persons all consider themselves professionals.

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

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

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 mistakenly focuses on two information items rather than one. If this is not caught, we invariably end up with overlapping response categories. For example, we wanted to know how former DOD employees had learned about postemployment restrictions. The word "how" in this context has various meanings: from a coworker, at a retiree meeting, from magazines or newsletters, during an exit interview at DOD, 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 may need to ask two questions: "From whom did you learn about . . . ?" and "Where were you when you learned about it?"

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 of 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 design relevant and appropriate items, the wording should be tailored to our respondents. An illustration is in the way we use medical terms, which depends on the background of most of the respondents. If we need to measure the receipt of health services from a nonmedically trained group, we use simple terms, give examples, and include identifying initials. Responses to

"What services have you received from your health maintenance organization in the past year?" might include

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: ophthalmology
4. ☐ Ear, nose, and throat care: ENT
5. ☐ Mental health or psychiatric service
6. ☐ Arthritis or rheumatism treatment
7. ☐ Allergy treatment
8. ☐ Immunization

USING CATEGORIES OF APPROPRIATE SPECIFICITY

The categories we use should be neither too broad nor too specific for our measurement purpose; the specificity should be tailored to each respondent group. 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 assessment. We do not need more specific information on the clarity of the delivery through diction, accent, and syntax.

A survey on water pollution further illustrates this point. When EPA asked paper-manufacturing plants about the acidity and alkalinity (pH) of waste water released into rivers, the response categories were not precise enough. EPA 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. In addition to losing the neutral measures, the categories were not mutually exclusive; scores between 6.5 and 7.5 could fall in two categories, 6 to 7 or 7 to 8.

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, respondents may feel they have misunderstood the question and return to study it again.

For example, in a survey on the attitudes of field staff toward relocating in Washington, employees were asked to rate the benefits and problems from the agency's perspective and from their own. The first option listed under personal concerns was "monetary loss." If "degradation in job efficiency" had occurred first, respondents might have wondered about the question's intent and the frame of reference.

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. Also, when each of the response categories is to be rated (for example, by degree of importance), subgrouping probably aids respondents in assessing each entry's relative value. Long response lists are also more subject to primacy effects; that is, if respondents are asked to select one entry from a long list, they tend to select one of the first entries.

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 long list, they will probably check something. 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 9

MINIMIZING QUESTION BIAS AND MEMORY ERRORS

By properly constructing our questionnaires, we can reduce the number of inaccurate responses from

- biased questions (questions that may influence the respondent to give information that differs from the true state of affairs) and
- memory errors (errors caused by a respondent's either forgetting that an event occurred or incorrectly recalling it).

QUESTION BIAS

Bias can occur in either the question or the structure in which the response must be given. Information from biased items is usually unusable, because the analyst cannot determine how or to what extent the information is distorted. Respondents may be

- unaware of the bias and respond in a way suggested by the wording,
- aware of the bias and deliberately answer in a way that does not reflect their true position, or
- aware of the bias and refuse to answer the question.

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, DOD 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 DOD, or did you leave without getting this information?"

Better presentations of these questions would delete status quo information.

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.

Bias from specific words

Certain words are "loaded" because they evoke strong emotional feelings. In our culture, such terms as "American," "freedom," and "equality" tend to evoke positive feelings and "communist," "socialist," and "bureaucracy" 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. For example,

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?

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 more.

Another example is, "Do you agree with radical black leaders that more members of their race should be hired by the building trade unions?" As the authors point out, such phrases as "people who don't work" and "radical" do not contribute to an objective frame of reference.¹

Another example, from a GAO study, involves a mail survey of private industry's views on competitive bidding practices for major DOD 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 overreport product cost and underreport product performance and delivery schedules on the theory that a firm will look better after having produced a product that costs less, performs better, and is delivered quicker than initially projected. Another approach would be to make a realistic bid specifying product costs, performance, and delivery schedules as you actually think they will be. Still a third approach would be to be optimistic in your bidding, on the theory that if you don't provide an optimistic proposal, you won't get the job. The question is, Which strategy gives the best probability of winning a federal contract: making conservative estimates, making realistic estimates, or making 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 report broken glass if we ask them what happened when the car was "smashed" but will not report broken glass if we ask them what happened when the car was "hit," even if the film does not show any glass breaking.

Unbalanced questions or presenting only one side of the argument

The wording of an item 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" ("satisfied or not") or word opposites ("support/oppose," "approve/disapprove"). Better wording for the questions

¹For further illustrations, see Warwick and Lininger, 1975 (the complete citation is in appendix I).

above would be, "Do you support or oppose the establishment of group homes for the mentally retarded in single family zones?" and "How satisfied or not are you with the training you receive?"

In constructing a question, it is very 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. It is possible that the distinction between "not satisfied" and "dissatisfied" applies to areas other than employment. In some instances, selection of appropriately opposite terms can be difficult. Pretesting question wording can ensure that both sides are properly balanced.

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 while others might judge satisfaction on promotion policies and types of tasks and responsibilities. Many times, 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 (1) the respondent has a vested interest in the subject and (2) we ask complex questions containing several aspects.

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, suppose we ask farmers if they should receive special agricultural weather reports free of charge. The question might be formally balanced by asking, "Do you think the government should provide agricultural weather reports free of charge to farmers or not?" However, we would expect farmers to favor such services. To obtain a more accurate measurement, we need to help respondents consider the question from a variety of viewpoints. For example,

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

In a survey question mentioned previously, program managers of major weapon systems were asked whether civilians or military should be program managers. Most of the respondents were military experts on the subject. To obtain opinions based on all aspects of the issue, we presented the pros and cons:

"A persistent issue is whether or not the PM position should be held exclusively by military personnel, exclusively by civilian personnel, or by both. There are advantages and disadvantages attributed to both the military PM and the civilian PM. Pro-military arguments claim knowledge and appreciation of the system (conditions, personnel, organization, etc.) and advantages of service affiliation. On the other hand, the military PM system is sometimes criticized for short tenure, valuing performance over cost, constraints on independent action due to the military rank hierarchy, and service/mission suboptimization. The question is, What should the military/civilian composition of the federal PM work force be?"

Broad questions contain many different aspects to be evaluated. For a variety of reasons, 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 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 occur when we have a short list of simple or straightforward responses or a very long list of short alternatives. In the latter, the list may be so long that the respondent becomes bored or fatigued and checks alternatives presented first. Recency effects seem to be more likely when we have lengthy or complex responses.

Although response-order effects do not always occur, they can cause serious bias problems when they do. We usually apply techniques to control or avoid their occurrence in two instances. First, when using scaled alternatives (such as five choices ranging from very satisfied through very dissatisfied), we present the least likely choice first. This counteracts primacy effects. For example, community opposition is frequently cited as an obstacle to locating group homes in residential areas. In surveying people who operate group homes for the mentally retarded and emotionally ill, we asked them to answer the following question:

Consider the individuals and groups in your community who were contacted. Overall, how did their support and opposition compare?

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

Second, in presenting a list with many response alternatives, we divide the list up into meaningful, shorter subgroups. The purpose of this technique is twofold: to counteract a primacy effect and to make the respondents' task as easy as possible. In asking people why they went into bankruptcy, we used the following technique:

EMPLOYMENT

1. ☐ Unable to work due to illness or accident
2. ☐ Period(s) of unemployment due to job layoffs, job changes, strikes, seasonal factors, etc.
3. ☐ Cutback in hours worked per week (e.g., loss of overtime; work slowdown; or, if 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 payments

The questions went similarly through two other categories for a total list of 16 entries.

Presenting choices in a logical sequence

In presenting a list of unscaled response alternatives (reasons for going bankrupt, characteristics of grazing land, and the like), we must put them in a logical order. That is, the options that are of primary significance to the topic being considered should be listed first. Because the list provides a frame of reference, commencing with arguments of lesser importance will deflect the respondent's grasp of the question's focus.

An example may clarify this. A questionnaire asked people why they dropped their memberships in health maintenance

organizations. Some primary reasons would include the ability to choose doctors and the quality of care. Paperwork and hospital decor would probably play lesser roles, and beginning the list of responses with these would be logically inappropriate and might confuse the respondent's understanding of the data we want.

Use of the "other" category and incomplete lists

As mentioned in chapter 8, we usually 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 give salient answers that we missed and decreases the item nonresponse rate. People who have difficulty making a choice will check "other" and write in an answer. This category also decreases overreporting that might occur in the categories that are listed. Rather than not respond, people tend to fit their answers into one of the available categories.

The factors that influence answers in multiple choice questions are very complicated. Many factors lead to both underreporting and overreporting. The net effect is underreporting. In general, the reasons reported are minimal estimates, especially in the "other" category.

It is essential that we analyze the "other" category to (1) determine the adequacy of the choices listed, (2) make adjustments for underreporting in the major categories (for example, one respondent wrote "availability of housing" under "other" when housing stock was listed as an entry), and (3) assess the complicated underreporting and overreporting tendencies.

Biased examples

Sometimes a questionnaire writer will provide examples to illustrate the kind of information needed. Examples should convey the range of information wanted. Single illustrations may cause a respondent to restrict a frame of reference. If we ask students how satisfied they are, if at all, with their training and mention the name of only one teacher as an example, we may get their evaluation of that teacher's class rather than of their training in general.

MEMORY ERRORS

Depending on the time since an event occurred and its saliency, memory error can result in either underreporting or overreporting. Memory errors reveal themselves in three ways: omissions (forgetting that an event occurred), intrusions (recalling an event that never occurred), and event displacement (miscalculating when an event occurred). Deliberate failure to mention an event, evasion, misinformation, and distortion cause

measurement errors, but these errors are not caused by memory failures. Since the deliberately false response is more likely to occur with highly sensitive or threatening topics or with topics that ask for self-evaluation, these problems are discussed in more detail in chapter 10, and so are similarly deliberate nonresponses.

Although research on memory has been conducted for some 500 years, the importance of memory error became readily apparent only with national and regional surveys, such as those conducted to study victimization, housing, and medical care. Comparative and experimental studies of the specific role of short-term memory in reading questionnaires are even more recent. Some studies used personal interviews rather than mail questionnaires; many used short questionnaires focused on broad attitudinal questions. GAO questionnaires tend to be long and ask for facts or expert opinions. Below, we present some general principles that reduce forgetting and increase the respondent's ability to accurately place events.

Using cues to help respondents retrieve data

Explicitly phrased questions that avoid special terms and difficult words, provide examples, state the exact range of information we are asking for, and provide a clear time reference help respondents search for the right memories. Providing a list of response alternatives further aids recall:

Which, if any, of the following community resources and services generally used by most of the group home residents are within walking distance (about a mile) of the facility? (Check all that apply. Check box 1 if none are within walking distance.)

1. ☐ No community resources or services within walking distance
2. ☐ Medical services
3. ☐ Social services
4. ☐ Drug store(s)
5. ☐ Food store(s)
6. ☐ Fast food service(s) or restaurant(s)
7. ☐ Variety or department store(s)
8. ☐ Movies
9. ☐ Library

10. ☐ Recreational facilities for children
11. ☐ Recreational facilities for teenagers
12. ☐ Recreational facilities for adults
13. ☐ Other (specify)

Cues are also provided in the following example:

"Rating only the person you relied on most, overall, how satisfied, if at all, are you with the way the individual handled your bankruptcy case? (Consider amount of time spent on case and clearness and completeness of the information you got.)"

In still another instance, cues were used to help reduce memory loss in crime reporting. Researchers found that dealing with the conditions and activities that the victims might have experienced produced more accurate recall, as in "Think about the times you came home late at night last year. Were you ever robbed or assaulted?" (Biderman, 1972)

Longer questions may help jog the respondent's memory

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 longer but may reap rewards by increasing a respondent's memory and comprehension.

Difficulties in remembering details

People remember essentials or the gist of an event better than its details. If we need highly detailed information, we should consider using data collection sources, such as observation, diaries, and records rather than self-administered questionnaires. Respondents to mail questionnaires can be asked to refer to their records; however, the burden of this may decrease response rates.

Recalling salient versus repetitive events

Events that are highly salient to an individual are more memorable than repetitive events. Salient events have been defined as events that are unusual or have economic and social costs or benefits; 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 we have a high level of

recall. People seem to lose their ability to distinguish between repetitive events, although they can recall or infer typical characteristics. For example, health-maintenance organization patients may have difficulty recalling the number of outpatient visits they made during a 6-month period but will remember or infer their mode of transportation, how busy the waiting room usually was, and how long it usually took to see the doctor.

Sometimes, measurements of routine or repetitive tasks cannot be made from self-administered questionnaires because recall is faulty. For example, GAO wanted to test the theory that hospital nurses spend about 10 percent more time delivering a routine procedure to the elderly than delivering it to other adult groups. Pretesting the questionnaire showed that nurses were able to accurately recall the types of tasks they did and to estimate the time they spent with the elderly for unusual nursing duties. However, for most of their routine tasks, the nurses could not estimate precisely enough the additional time they spent with the elderly. The evaluators had to use on-site observation instead of a self-administered survey.

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"). For questions about the frequency or timing of salient events, respondents should be asked to report on events that occurred during the last year. Omissions will be few and telescoping will be minimized.

For highly repetitive and habitual events, however, it is probably best to ask respondents to recall data from 2 weeks to 1 month earlier. For events of intermediate saliency, about 3 months gives optimal accuracy. These time periods seem to provide the best conditions for balancing omissions caused by forgetting and errors caused by incorrectly remembering an event.

In summary, questionnaire designers must be very careful to account for the effects of memory in reporting. They must consider the short-term memory bias introduced by position, emphasis, and complexity or by the simplicity of the preceding text or succeeding answers. The questions must be structured and written in ways that aid recall. 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 must know the limitations of a respondent's memory--that is, the types of events and time periods for which recall is usually very poor.

CHAPTER 10

MINIMIZING RESPONDENT BIASES

The previous chapter discussed the response inaccuracies that can occur when we ask biased questions. Bias can also occur in the responses to our questions because of

- a respondent's style in question answering, such as the tendency to agree regardless of the issue, and
- topics that respondents perceive as highly sensitive, objectionable, or threatening.

This chapter discusses 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. Most of the research on response styles has occurred in the fields of psychology and sociology through applications in testing and surveys. 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.

Response styles can interact to compound a distortion. For example, the tendency to underreport behavior that is socially undesirable (for example, drinking liquor) may be further exaggerated if the behavior is presented in the extreme. This is because people are reluctant to report behavior that is not socially desirable and they are also reluctant to report extremes for most behavior. Hence, question writers must be aware of response-style distortions and the ways to account for or counterbalance them.

Socially desirable responses

Respondents may select socially desirable answers over other choices. Socially desirable responses represent culturally accepted norms for opinions and behavior (for example, compliance with regulations and belief in "truth, justice, and the American way").

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 require substantially more responsibility. To reduce the potential for overstatement, the questionnaire should describe jobs or give examples tailored specifically to the respondents.

To reduce overreporting or overstatement of socially desirable responses,

1. Ask specific questions. Broad questions permit the respondent to "truthfully" check a socially desirable response option by focusing on a small number of factors. For example, the shipping clerk can check "traffic manager" if answers to detailed questions about job responsibilities are not required.
2. Make a single question containing a socially desirable response into two or more items. The extra items measuring behavior can be used, because respondents are more likely to answer truthfully about verifiable behavior. Also, a series of questions can provide a respondent a "face-saving" escape. Although the behavioral question may not permit the respondent to register a socially acceptable 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 time it would take them to prepare to do the tests:

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 by answering questions in ways that place them in the best possible light. 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. Assuming a bias toward making a good impression, it is likely that "took on too many debts at one time" would be underreported as a significant factor and "credit was too easy to get" would be overreported. To overcome this tendency, we made sure these answers were not placed in the most prominent positions but were listed midway in a checklist of several other plausible choices in a matter-of-fact manner. This approach helped the respondents place response options in a more objective frame of reference. Also, "took on too many debts at one time" and "credit was too easy to get" were actually two sides of the same coin. An analysis of the responses from both items provided a closer approximation of the number of people who overextended themselves than either item alone would have provided.

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, they tend not to use polar positions. If we present people with three choices (for, neutral, and against, for example), they tend to select the middle category. To counteract this tendency, we 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, we avoid asking people how much they agree or disagree with a particular item. Besides offering the opportunity for a "yeasaying" bias, asking only for a respondent's agreement or disagreement provides very limited information. For instance, in a study on federal grants, the study authors attempted to assess users' perceptions of specific grants programs. They used a "Likert" or agree-disagree question format, as we show on the following page:

Please indicate the extent to which you agree or disagree with each of the following statements. Do this by writing strongly agree, generally agree, agree as much as disagree, generally disagree, or strongly disagree in the space after each statement listed below.

1. Federal funds are not necessary_____
 2. Federal funds take too long to obtain_____
 3. You need more funds than the federal government can grant or loan to you_____
 4. Other funding concerns_____
- (Please specify)_____

Positively worded questions like this may elicit a yeasaying bias and not probe or provide much information as, for example, "Why aren't federal funds needed?" or "Why does it take too long to obtain funding?"

HIGHLY SENSITIVE ITEMS

As mentioned in chapter 6, highly sensitive questions should be written with care and used only when the information is vital to the evaluation or audit and cannot be otherwise obtained. Unless they are carefully handled, response refusals and other report inaccuracies occur.

Highly personal items, such as data on income, sex, marital status, education, and race, may be perceived by respondents as intrusive and perhaps objectionable. Personal questions should be included only if they are 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 respondents' 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 not be used in GAO studies.

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 worth while? And how will the sensitive question affect GAO's image among respondents and the public? If we decide that sensitive items are necessary, the

following guidelines should be used to reduce underreporting and answer bias.

1. Explain to the respondent our 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, makes the behavior in question appear socially acceptable.
4. Present the request in as matter-of-fact a style as possible.
5. Guarantee confidentiality or anonymity, if possible.
6. Make sure the respondent knows we do not plan to use the information in a threatening way.
7. Explain how the information will be handled.
8. Avoid crossclassification that would pinpoint the answers.

For example, when we 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:

During the year that 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 we need to find

out about the job-hunting activities of the unemployed. The question series might be developed like the one below:

1. Have you had any difficulty looking for work? (Check one.)
 1. ☐ Yes
 2. ☐ No
2. If yes, which of the following factors have caused difficulties in looking for work?
 1. ☐ Illness
 2. ☐ Lack of transportation
 3. ☐ No child care arrangements
[etc.]
3. In spite of these difficulties, in the last month, how many contacts were you able to make?

_____ (number of job-related contacts)

Notice that item 1 recognizes that looking for work is difficult. This is to put the respondent at ease and reduce the threat of revealing private, possibly embarrassing information. The information asked for in question 2 was not needed; the question was asked because it followed logically from question 1, designed to make not looking for work socially acceptable by providing very good reasons. Question 3 focuses on the real data needed, but without questions 1 and 2 the answers would be overreported. Note that question 3 is also phrased to put the respondent at ease, even if few or no job-hunting attempts were made.

Using a specified time reference can reduce a question's threat. For instance, if we need to find out whether people are coming in late for work, we 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 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 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 9 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?"

CHAPTER 11

MEASUREMENT ERROR AND MEASUREMENT SCALES

Answers to our questionnaires provide quantitative data. For example, they tell us how much, how often, how big, how adequate, how better, or how worse. The problem is that they are not exact measuring instruments. But we can often translate respondents' reports into reasonable measures if we understand something about measurement error and measurement scales.

MEASUREMENT ERROR

In several of the preceding chapters, we have discussed situations in which mail questionnaires are not appropriate because people are not knowledgeable, accurate, or honest reporters. We also explained when, where, and how to use questionnaires to obtain reasonable measures. But we did not discuss the errors we make when we obtain these measures. These are called "measurement errors" and are an important consideration, because measurement error contributes, along with errors from other sources such as sampling and data reduction, to the total error in the results we report.

Virtually all measurements contain some amount of error, and this is especially true when we are trying to acquire information from people. One of the principal purposes for using questionnaires and structured interviews is to keep measurement error within acceptable limits. Unstructured approaches to data collection, such as informal interviewing, tend to have much greater potential for measurement error and, consequently, provide a weaker basis for our results.

When we use a data collection instrument such as a questionnaire, we are typically attempting to acquire information about a person, a thing, or an organization. For example, we may want to know a person's annual income, so we ask a question such as, "What was your total personal income, last year?" We have in mind that there is a true value for this figure, and we want the person to respond with the true value. In general, however, the respondent will not give the true value, and the amount of discrepancy is called a "measurement error."

Why would there be a measurement error? Perhaps because of the way the question is stated. Or because the preceding sequence of questions led the respondent to an interpretation of the current question different from that which we had intended. Or because of the respondent's perception of how we might use our knowledge about true income. A number of factors may influence the response, and much of the advice in this paper is directed at trying to minimize measurement error.

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 fixed amount. For example, the way we phrase a questionnaire item about income may cause respondents to fail to include a particular category of income. Another example is that some employment surveys consistently overestimate the real level of unemployment because of the way that their 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." A measurement process may lead respondents to underreport or overreport inconsistently by a variable amount. 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 ambiguous, there is much opportunity for random error.

Measurement error exists in all forms of empirical inquiry. Whether the discipline is auditing, evaluation, physics, or any other, investigators must contend with error in their measuring instruments. There are three main ways of doing this: (1) constructing the measuring instruments and using them in ways that make error small, (2) estimating the size of errors from bias and making corresponding adjustments to the results, and (3) estimating the range of random errors and reporting this information. Procedures 2 and 3 are beyond the scope of this paper and will be covered in forthcoming PEMD transfer papers on the basics of data analysis and measurement.

However, one general point should be stressed here. The first procedure has to do with controlling error by following good practice in the development and use of questionnaires. Errors tend to be minimized when the measuring process is standardized: when the observers and the equipment operate in the same way under the same conditions. Standardization leads to gain, whether we speak of the physicist's thermometer or the evaluator's questionnaire. For mail questionnaires, every respondent reading the form should interpret each of the questions the same way. This is why the preceding chapters emphasized the need for structure and the need to consider the effects of format, appropriateness, qualifications, clarity, memory, and respondent bias. Standardization is achieved by adhering to the principles and practices of good questionnaire design and, then, by carefully testing the instrument.

MEASUREMENT SCALES

In chapter 5, we discussed how different formats permit different levels of measurement. These levels, called "scales," determine the types of analysis technique we can use. Higher scales give us more information and they allow us to take advantage of the more powerful statistical techniques. Hence, in selecting a question format, we think ahead to the point at which we will have finished our data collection and will be starting our analysis. That is, we try to use the scale that will let us use the preferred statistical techniques without prohibitively increasing costs or respondent burden.

Although several sets of scales have been determined, the best-known set of scales is probably these four: nominal, ordinal, interval, and ratio. A scale is a set of categories by which we differentiate measurements.

Nominal scales

Nominal scales allow us to categorize measurements without placing the categories in numerical order. For example, a person is either male or female; no numerical order is implied by the classification. Other examples are states, colleges, east coast and west coast, shippers, political affiliations, and marital status. Most questions that are answered by checking one or multiple-choice responses involve nominal measures.

The requirements for a nominal category are relatively simple. The categories must be mutually exclusive, and we must be able to place each case either in or out of each category. For example, enlisted personnel were asked whether they received mission and military occupational specialty (MOS) training. Respondents could be classified as having received only mission training, only MOS training, or both kinds of training.

Since nominal scales allow for only the simplest of statistics, we are limited in the ways we can describe, use, analyze, and interpret the data. Frequency distributions, modes, crosstabulations, and several measures of association (or statistics that show the direction or magnitude of the relationship between two variables) are used most commonly with nominal data. A crosstabulation shows, for instance, the proportion of people who fall into two overlapping variable categories. In the military example above, it would be the proportion who received both mission and occupational training. A frequency distribution tells us the number of persons in the population who belong to the category and the number who do not. The mode is the category that most people are in. And the percentage is the percentage of the total population in a category. We can estimate the statistical certainty with which persons were assigned to a category, and we can estimate the certainty that categories have the same number of members or not.

Ordinal scales

Ordinal measures rank people or things. We can know not only whether a person or an object falls into certain categories but also the numerical order of the categories. Although the order of the categories shows the relative position of things, for example, it does not tell us the extent of the difference between the positions, as in a classification of the people by the level of school they last attended (grade school, high school, college, and graduate school).

What we do know is that the categories lie along a dimension and that an observation placed in category 1 will always be greater than an observation placed in category 2, an observation will be greater in category 2 than in category 3, and so on. We have enough knowledge to calculate a median. (A median is the point on the measurement scale at which half the observations are below and half are above.) We also have a gross measure of dispersion (that is, the degree to which the measures are spread out or clustered). And we have some measures of association. For example, suppose we categorize persons on two dimensions: (1) young or old and (2) low or high medical expenses. From data collected on such measures, we could make a statement such as "the older people get, the more likely they are to have high medical expenses."

Interval scales

As with ordinal scales, the categories in interval scales fall along a dimension in order. The important difference is that all the categories are equal distances from one another. That is, the range between the highest and lowest scores is divided into a number of equal units, like a thermometer. Thus, we can tell how far apart the first measure is from the second and how close each measure is to the low, middle, and high ends of the scale.

Interval scales give us a lot of power, because they are metric; that is, we can add, subtract, multiply, and divide measurements on such a scale. We know that the difference between 20 degrees and 30 degrees is the same as the difference between 50 degrees and 60 degrees. We can calculate averages, medians, standard deviations (a powerful measure of dispersion), frequencies, and modes. We can use most of the more powerful and versatile statistical analysis techniques, such as correlational analysis, analysis of variance, and t tests, to tell us whether variables are associated and whether the means of two groups are different. We can describe the distribution of observations and see how the observations are clustered or dispersed along the complete range of possible values. We can look at the effects of several combinations of variables all interacting together. Therefore, for most of our questionnaire formats, we consider ourselves very lucky if we have an interval scale.

Interval scales have one problem, however: they have no true zero point. We cannot say that 60 degrees Fahrenheit is twice as hot as 30 degrees Fahrenheit, because zero is not the extreme end of the Fahrenheit scale. That is, we cannot get to a point at which there is no heat on this scale. Zero degrees does not mean no heat; it is just another point on the scale. So when we say that the temperature is 30° F or 60° F, all we are saying is that the temperature is 30° or 60° hotter than zero. We can say that 60° F is 30° F hotter than 30° F, but we cannot say that it is twice as hot as 30° F.

Ratio scales

Ratio scales are interval scales with a true zero point. Money, for example, is quantified by a ratio scale. We can have no money. The difference between \$50 and \$100 is the same as that between \$150 and \$200, and \$200 is twice as great as \$100. Among scales, ratio scales provide us with the most complete descriptive information. Also, we can use the broadest range of analysis techniques, including most of the statistical methods that apply to interval scales. Parametric and nonparametric techniques are appropriate for ratio scales. Most fill-in-the-blank questions involve ratio scale measures such as dollars, number of staff, and frequency of occurrence.

Equal-appearing intervals

Frequently, we make observations on a variable for which the scale naturally has many small categories but we choose to use a coarser scale. For instance, people might be reluctant to tell us their income (a fine scale), but they will tell us if their income falls into a certain broad category. When using this technique, we 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.

However, when respondents check the "\$15,000 up to \$20,000" category, we do not know if they made just a little more than \$15,000 or just a little less than \$20,000. Without certain assumptions, we cannot treat this as interval data, because we have a cross between an ordinal scale and an interval scale. When this happens, we have to learn more about the distribution. If we are reasonably sure that the actual incomes in the category are symmetrically distributed, we take the midpoint of this category as our data point. This assumption allows us to use some of the methods reserved for interval data. We would run into trouble with this assumption, however, if a disproportionate number of people made just over \$15,000 or just under \$20,000.

Another example of the connection between the questionnaire format and the measurement scale can be seen in the Likert questions discussed in chapter 5. The Likert format has five

broad categories. Should we consider it an interval scale, so that we can use the statistical techniques for interval data, or should we consider the format only ordinal? That is, we do not know that the categories (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree) have equal intervals. We must consider the literature, our own experience, and our data to make a judgment.

In Likert formats, we almost always treat the information as ordinal or ranking data. However, for some of the other intensity scales discussed in chapter 5, we can sometimes make a better case for an interval interpretation. For example, we may have some evidence that "generally satisfied" falls three quarters of the way between "very dissatisfied" and "very satisfied." However, even in such situations, we usually show the proportion in each group and consider the category information as ranking data.

Since rating categories treated this way do not give us 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 would have to assume that "somewhat difficult" falls one fourth of the way between "not difficult" and "extremely difficult."

However, such assumptions are very hard to make. When we make rating category scales, we are very careful to assign them on the basis of our knowledge of the variable in question, the literature, our past experience, and our pretest results. Sometimes we even conduct a special study to verify our assumptions. But if we are uncertain about the assumption, we treat the observations as ordinal data. If the assumptions are reasonable and the conditions are right, we sometimes treat attitude measures like "satisfaction" as interval data.

CHAPTER 12

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" Because most people are not accustomed to filling out questionnaires, we need to guide them through the experience. 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 ABOUT OUR LINE OF INQUIRY

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 GAO, 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, relevance, and protections of confidentiality or anonymity.
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 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. Therefore, we need to be aware of questions that respondents will perceive as relevant and important to the questionnaire's stated purpose and of the sequence that respondents will expect.

We should strive to present items 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 we follow the natural or chronological flow of a topic, we stand a better chance of helping respondents recognize and recall the information we need.

USING SUBTITLES AS CUES

Some say that mail questionnaires are disadvantageous because respondents can look ahead and see the types of questions that are to be answered. The notion is that perhaps having advance information will influence how one answers. But there is another side to the coin. Related items that are grouped and accompanied by subtitles help the respondents quickly grasp the scope and nature of our inquiry. With the necessary framework in mind, they probably provide more accurate and comprehensive answers. For example, in a GAO evaluation on how personal bankruptcy cases were handled under a federal bankruptcy law and on factors that led to personal bankruptcy, we used the following subtitles set off in bold capitalized type:

- BANKRUPTCY PROCEEDINGS
- CHOICE OF CHAPTER
- CHANGE OF CHAPTER
- STATUS OF CHAPTER 7
- EXEMPTIONS
- INFORMATION ON BANKRUPTCY
- BACKGROUND INFORMATION (on the respondents)
- COMMENTS

Individual items within subtitled groups should unfold in a meaningful fashion. For example, the questions under "bankruptcy proceedings" were

- "1. Under what name was the bankruptcy petition filed?
2. Who filed the court papers to start bankruptcy proceedings?
3. What individual did you rely on most for assistance in handling the case?
4. Did the individual discourage you about filing, encourage you to file, or neither?
5. How satisfied or not were you with the way the individual handled your case?

6. If dissatisfied, for what reason or reasons were you dissatisfied?"

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 we need their opinions in key areas. It should introduce the language and rules of the questionnaire. Potentially objectionable and threatening questions should be avoided as initial questions, because they may discourage recipients from completing the form.

If possible, the opening item should apply to all the respondents. Questions with such response options as "not certain" and "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. 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:

STOP.

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. PLEASE MAIL THE QUESTIONNAIRE AND THE POST CARD SEPARATELY.

We should avoid starting a questionnaire 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 our information needs and may not provide the range and depth of data we want.

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 design complexity.

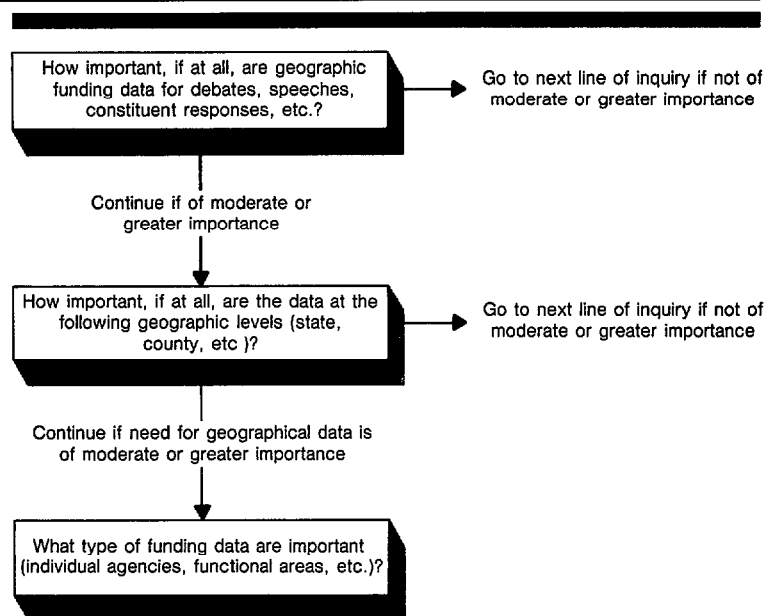
Demographic questions are usually placed near the end of a questionnaire, because they may be perceived as highly personal and as perhaps less important to the questionnaire's purpose.

The placement of demographic items depends, however, on the topic and the audience. For example, military personnel are accustomed to providing rank and grade, length in the service, and similar personal data. In a survey of practicing physicians at DOD hospitals, it was quite appropriate to ask them about their rank, grade, medical education, certification, and the like at the beginning of the questionnaire. If the demographic items seem less relevant to the questionnaire topic, we may want to explain their purpose to respondents in words such as "This information will help us in our analysis of responses."

OBTAINING COMPLEX DATA

Sometimes a mail questionnaire obtains very complex data. Because the form is self-administered, we need to design it 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 of not only geographical and funding categories but also each particular combination (such as program data at the state level). In total, we needed 288 separate answers. Figure 12.1 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 12.1: Sequence of Questions on Funding Data



USING TRANSITIONAL PHRASES

Sometimes the respondents' 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, one section of the questionnaire was concerned 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 on accountability.

Transitional phrases may be particularly necessary if a series of complex questions covers several pages. For example, in a survey of state coordinators of 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 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 "

PUTTING SPECIFIC QUESTIONS BEFORE OVERALL JUDGMENT QUESTIONS

The total collection of items in a questionnaire provides a context for interpreting the individual items. Specific questions that ask for facts and lead up to a more general question asking for judgment promote a common frame of reference, so that the respondents can base their judgments on the same set of factors. Proceeding from specific data to general judgments

also helps respondents recall and assemble the range of data needed for informed judgments. Two examples illustrate the specific-to-judgment rule.

A GAO study surveyed veterans concerned about their exposure to agent orange. They had sought assistance at medical centers operated by the Veterans Administration. A series of items asked about the types of examinations and laboratory tests given by the medical centers and the nature of the respondents' health problems. After asking for information in many specific areas, we asked respondents to rate their satisfaction or dissatisfaction with the medical help they had received. Because the overall judgment question followed specific items, respondents were better able to provide thoughtful answers.

In another survey, officers at financial organizations were asked about the Security and Exchange Commission's program on lost and stolen securities. Before focusing on their views of the effectiveness of program and proposed changes, we asked for information on about 50 items. Several judgment questions, some quite detailed and complex, followed.

DEALING WITH ADVERSE INTERACTIONS

Sometimes the interaction of items produces distorted data. Data can be distorted when we ask a general question in connection with a series of detailed questions that focus on only a portion of the general item. For example, managers of homes for runaway children might be asked about federally funded alcohol and drug abuse services but not about other federally funded programs for runaways, such as education, family, and mental health services. If we then asked a general question about federally funded programs for runaways, the respondents' answers would be likely to be unduly influenced by the drug and alcohol abuse program.

Item interaction can occur even though we give equal attention to various aspects of a topic. Inquiries that ask people to evaluate a topic from both a personal perspective and someone else's are probably very difficult for respondents to answer in a neutral fashion. 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 we define the scope of a general question. 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, businesses have a right to lock them out.

If a specific question precedes a general question on the same topic, respondents may "redefine" what the general question refers to. For example, answers to the general question, "Taken all together, how would you say things are these days--would you say you are very happy, pretty happy, or not too happy?" may change, depending on whether this question comes before or after the specific question, "Would you describe your marriage as very happy" One explanation for the endorsement differences is that people redefine what general happiness covers and either include or exclude marital happiness from consideration. The central point to note is that when we ask a series of opinion, attitude, or other judgmental questions, it is essential to study them for the potential interaction from sequencing.

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 the form, the respondents' 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 respondents' 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 respondents' effort and attention probably rise again. This accompanies a feeling of investment--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 13

QUALITY-ASSURANCE PROCEDURES

We check the quality of questionnaires 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, we pretest the questionnaire on selected persons who represent the range of conditions likely to influence the evaluation's results. Also, we usually send the questionnaire outside GAO for review by experts who are familiar with both the issue area and the respondent group.

Pretesting and expert review are very important in the development of a mail questionnaire. We need to ensure that the instrument will adequately communicate what we intend, 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. If we do not take these steps, we will almost always wind up with undetected design flaws and may also overlook critical factors in the evaluation. Undetected design flaws and incomplete measurements of critical variables are often serious enough to severely compromise the results of an evaluation.

In our quality-assurance effort, we frequently do validation or verification work, analyze nonresponses, and conduct reliability studies. We will briefly discuss these methods before discussing pretesting, expert review, and the role of validation, verification, nonresponse analysis, and reliability in quality assurance.

Validation is an effort to ensure that the questionnaire is actually measuring the variables it was designed to measure. For example, we may think we are measuring metric tons (a weight measure) of shipped cargo but then find out that our respondents are giving answers in shipping tons (a volume measure). Verification is a way of checking or testing our observations against the same kind of information from another, independent source. For example, we might check self-reports of visits to doctors against physicians' records.

Verification is different from validation. To validate, we must provide evidence that we are measuring what we say we are measuring. For example, if we are interested in the quality of health care, then the number of visits to doctors may not be a good 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. Hence, when we checked patients' self-reports against physicians' records, we would be testing the soundness of self-reports only as a measure of visits (verification), not as a measure of the quality of service (validation).

We sometimes test the reliability of questionnaire results by determining whether a question always gets the same results when repeated under similar conditions. Reliability is a measure of stability. It is different from verification and validation. Because reliability does not ensure verification and validity, we can have highly reliable answers that are not verified and are invalid.

We analyze item and questionnaire nonresponses also because high or disproportionate nonresponse rates can threaten the credibility and generalizability of our findings. This is important because, if only half the people respond, we do not know anything about the other half. The people we do not know about may be different from the others. Unless we can show that nonrespondents are not importantly different from respondents (an unlikely possibility) or that our nonresponse rate is small, our ability to project from the sample to the universe will be weakened.

Why do we have to validate, verify, and make reliability checks? We have to do much of this work because most of the time GAO cannot use instruments that have been already tested and developed. We are either measuring things that have not been measured before or measuring previously measured things under different circumstances. So we must do our own instrument development work. Nonresponse checks, however, are routine for all questionnaire work. Although GAO's procedures usually result in a high response rate, it is important to understand how nonrespondents may differ from respondents.

PRETESTING

By testing the questionnaire before we distribute it, we can assess whether we are asking the right group of people the right questions in the right way and whether they are willing and able to give us the information we need. Pretests are conducted with a small set of respondents from the universe 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.

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 we intended? Are the response choices appropriate and comprehensive? Should the question be more specific? Is the time frame

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 give us the needed information? Have we overlooked a critical construct or variable? Is the variable measured in sufficient detail?
4. Can and will the respondent give us the data we 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 objectionable or threatening? Does the questionnaire adequately motivate the respondent to provide us with 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 pretest procedures and about the questionnaire's content. However, if this is not possible, 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 actually 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 should be able to conduct the remaining pretests alone.

How do we select and contact pretest interviewees?

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 our 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--for example, 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 we 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, an interviewee should be selected whose language skills are somewhat less strong than those of the majority of potential respondents.

In principle, we would like to test enough people 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 candidates should be selected because they represent or have knowledge of the the range of characteristics or conditions we are likely to encounter--young and old, experienced and inexperienced, large and small companies, efficient and inefficient organizations, and so on. For example, in a survey of migrant workers, we pretested at the geographical beginning of the northward migration in Florida, Texas, and southern California and also at the middle and northernmost points, in order to catch the range of conditions of the different streams of migrant workers as they moved northward.

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. Arrangements are 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. It sometimes happens that the pretest subjects cannot be contacted by phone. This might happen 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 we communicate our request for pretesting, because some people react with discomfort to the word "test." This kind of reaction can be allayed if the evaluator explains that we need the interviewee's comments and criticism to test our 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; we are interested only in finding out how well our questionnaire works.

How do we conduct the pretest?

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

Introductory comments

The following points, which are mentioned in the telephone contact, are briefly covered again at the beginning of the pretest session. We should

- state the role of GAO;
- state the roles 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 was present; we should 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 we want the interviewee to look for (for example, an item may ask for dollar amounts by calendar year but 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 count all the times each patient left who also entered more than once during the year, or the list of options fails to include a critical component, or the interviewee is not sure of a particular response but no category (such as "Not sure") has been provided for

this, or a skip in the instructions is confusing, and so on);

--tell the interviewee that we will be following the sequence of questions on our own copy of the form in order to monitor the flow of questions, thus addressing any potential concern interviewees may encounter who notice that we are entering information on our form;

--state that we appreciate frank and honest answers and thank the interviewee for assisting us.

Completing the questionnaire

The pretest administrator carries out five tasks while the interviewee completes the form.

Record the time it takes to complete each item. At the beginning of the pretest, we should position ourselves so we have a clear view of the respondent's questionnaire and face. The start time is recorded at the top of our form. As the respondent works, we count silently the number of seconds it takes the interviewee to read the instructions or complete a question, and we record this time next to the relevant section on our copy of the form. We try to be unobtrusive. If the interviewee asks a question or the test is otherwise interrupted, we note the time taken out for the relevant item. 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.

Record questions asked and clarifications made. When the interviewee asks a question, we record key words or verbatim text, as well as our response, next to the relevant item. These comments are used as an aid in debriefing and in item rewriting. If the interviewee is confused about what a question means, we provide a straightforward answer. Probing should be done during debriefing, rather than during the test, to see what the problem was. We should pay particular attention to how we answer any questions the interviewee raises, and we should be careful when providing explanations or alternative wording. In deviating from the prescribed text, we may rephrase questions and bias the interviewee toward a particular response.

Note nonverbal behavior. We 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. We can use nonverbal observations as signals for questions we should ask during debriefing.

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

Note erasures, uncompleted items, errors, and inconsistencies. These types of responses may indicate questionnaire design flaws. We can pick these up as we review the respondent'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 respondent's answers and the GAO staff'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.

We begin the debriefing by stating its purpose and telling the respondents that we will be drawing on their experiences and judgments to

- ensure that the intent of each item is clearly conveyed,
- evaluate the relevancy of items, and
- identify item-design deficiencies.

We review in detail the respondents' questionnaires and get their feedback to our probing. The major problems to look for are

1. improper question format,
2. inappropriate questions,
3. improperly qualified questions,
4. inappropriate language,
5. failure to present an inclusive range of mutually exclusive alternatives,
6. complex questions,
7. unclear questions,
8. question bias, and
9. improper scales.

In discussing questionnaire items, we 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, indicating that the respondent missed certain key considerations; (6) questions the respondent 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 interviewer's approach in debriefing is nondirective. We try to elicit the interviewee's comments, problems, and reactions to the questionnaire without leading. We use general comments to get the interviewee to reconstruct the questionnaire experience. For example, we use the respondent's answers or our observations of behavior 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 we let the respondent tell us the reasons behind the behavior.

Some areas may need a more direct approach. If "don't know" is the answer supplied, we can probe to see whether the respondent is being evasive. If we believe the respondent has an answer, we can push a little. But we should not push so much that a true "don't know" becomes a bad response.

When the debriefing has been completed, we thank the interviewee for helping us perfect the questionnaire. As soon as possible, our comments and observations about the pretest should be recorded.

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 whose program is under review serves in this capacity. By obtaining agency input at this stage, we avoid problems after data collection, when time and money have been spent.

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.

VERIFICATION

Developing items that are free of design flaws is one part of a larger effort to ensure the credibility of the questionnaire data. Another part is verifying the answers. As we stated in the beginning of the chapter, verification consists of comparing our observations to those obtained from an independent source of information. For the most part, this work is simply to determine whether we can trust self-reports. Usually we do this by comparing a respondent's questionnaire reports with evidence developed from an "on-site inspection."

For example, the evaluators may compare the respondent's assessment of the quality of their houses with the evaluator's on-site inspection reports. We might check self-reports of attendance against time sheets, self-reported organizational affiliations against the legal documentation authorizing affiliations, self-reported settings of wall thermostats in homes against the actual setting of the thermostats, or self-reported company policies against a company's published policy manual.

These verifications are usually conducted on a statistical sample of the respondent population. However, verification is sometimes conducted with a judgment sample considered typical of the population.

There are other ways to verify the results. For instance, we can get confirmation from other sources by showing that other studies demonstrated similar findings or by crosschecking the results for internal consistency. In one study, we compared the reported population characteristics against the "fourth-count census" report for the area in question. In another case, we compared the results of an employee job-activity survey with results reported by a consultant study and a survey of supervisors.

We verify also by including consistency checks in the questionnaire. We might ask how much time it takes an employee to do a task and the number of times it is done each week. This could be crosschecked with another question that asks for the percentage of total time the employee spent doing this activity.

VALIDATION

Validation is usually different from and more important than verification. We want to show that the observation measures what it is supposed to measure. To illustrate, we can consider the following case. A GAO study planned to use the accounting

definition of profit (income minus the sum of expenses and capital depreciation) to assess economic return. Case studies of a sample of firms showed this concept to work well for big companies but not for owner-operators or proprietors. This was because the owner-operators did not draw a salary or pay themselves interest on the money they loaned their businesses. For big companies, these were considered expenses. Therefore, big companies had a much higher economic return than small companies for the same profit margin. For this and other reasons, profit was not a valid measure of economic return for many organizations in the sample.

The best way to demonstrate validity is to demonstrate the relationship between the measurement and the construct we seek to measure in a setting as controlled possible. We call this "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 respondents' ratings of the difficulty of the items. As the difficulty of the items increased, so did the mistakes, the respondents' ratings of difficulty, and the response times. We concluded that time-to-complete an item could be taken as a measure of the item's difficulty.

In another study, evaluators used supervisory ratings as a measure of performance. To validate this, they had supervisors rate employees as they completed a number of lengthy performance tests. The evaluators then compared the supervisors' ratings with performance test scores to check the validity of the ratings.

Few measures are completely valid, so the more rigorous and varied the validity tests, the stronger the case we make for our measure. There are a number of other ways to test validity. Although most of them are less convincing than construct validation, the method discussed above, 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 what we call "content validity." In this approach, we might ask experts to make sure that the measure includes the content we 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).

We can also test validity 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, our confidence in the validity of the measure is strengthened.

Although the rigor and pluralism of methods that are used determine the credibility of our claim for validity, there is a limit to our resources. We tend to validate most often when the measures are complicated and abstract or unproven or critical to the study findings or likely to be challenged. We may verify but not validate measures that are self-evident, uncomplicated, or concrete (sex, staffing levels, and budget size) or that have a long history of successful use.

Furthermore, verification can sometimes serve as validation, if the verifying agent or instrument can function as a validating criterion. For example, hospital records may serve to both verify and validate the measures of the occurrence and severity of reported accidents.

ANALYSIS OF QUESTIONNAIRE NONRESPONSES

In mail surveys, we rarely get an answer from every questionnaire recipient. One reason is nondeliverable mail. Recipients who do not meet the selection or eligibility criteria, people who move out of the universe, and death are also factors. Thus, our sample of recipients will shrink somewhat when we consider the nondeliverables and the ineligible. Some people receive the questionnaire but choose not to answer it. For projects in which we are seeking to generalize from the sample to the universe, our not getting an answer from everybody in the sample threatens the representativeness of the sample.

When people select themselves out, the sample is no longer random. This compromises our ability to generalize to the

universe. If the nonresponse rate and the nondeliverable rate are high, so is the threat to generalizability.

The nonresponse rate is calculated by using as a base the number of people in our sample who were eligible to answer the form. Those who do not meet the criteria, nondeliverables, and death are excluded. (We always report the nondeliverables and analyze this group for causes that could reflect undersampling.) The nonresponse rate is the number of people who received the questionnaire and were eligible to answer divided into the actual number who did answer.

The response rate should usually be at least 75 percent (a standard used by most practitioners; small to moderate differences between the respondent and nonrespondent populations 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.

Unless the response rate is very high (over 95 percent), the nonrespondent population should be analyzed. A comparison of respondents and nonrespondents with regard to demographic and other important characteristics will 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). If the nonrespondents had differed from the respondents, the accuracy of our results would have been seriously compromised.

The work papers should analyze the composition of the nonrespondents, indicate the number and type of categories excluded from our expected universe or sample, and document our attempts to verify or trace the correct addresses of those who could not be reached by mail. Also, if a nonresponse bias is detected, the survey results should be adjusted. For example, if a disproportionate number of nonrespondents are from California and 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 a random sample of nonrespondents to obtain answers to key questions or to find out why they did not complete the form. This information helps us assess the data that were returned. A discussion of the nonresponses should be included in the work papers and in the discussion of methodology.

In addition to the people who do not answer the form, some proportion of the people who answer the form do not complete some items. The average nonresponse rate should be calculated for each item and added to the survey nonresponse rate, 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, we exclude the item from our analysis, the final report should disclose this. The item nonresponse analyses should be included in the work papers and the discussion of methodology.

TESTING RELIABILITY

"Reliability" refers to the consistency of measures. That is, a reliable measure is one that, used repeatedly in order to make observations, would tend to produce the same result every time. Not, perhaps, the same every time but with consistency. Some measures may be extremely unreliable.

Testing reliability is difficult, expensive, and usually low in priority for GAO reviews. It is difficult and expensive, because we have to either replicate the data collection or return to those we questioned before. People do not like to be retested. It is not a high priority, because the variables we measure are relatively stable for the time periods in question. We are, however, careful to pretest this assumption.

In some situations, we do need to test the reliability of the questionnaire answers. First, respondents as a group may lack motivation or interest. Given this situation, they may not invest much care or thought in the questionnaire and their answers may vary over time. Second, if we expect respondents to exaggerate, a retest using the same questionnaire may give us better data and a more accurate and precise answer. This is because, on the second time around, a respondent is likely to focus more attention on the issue and may be more careful with

the answers than during the first test. 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 may produce different answers if it is completed while reservists are at summer training with their units. Fourth, if we anticipate that most respondents will take an extreme position toward the area of investigation, we should retest them, because extreme values are sometimes subject to change.

It is important to note that the procedures for testing the reliability of answers are different from those for validating answers. When we validate information, we 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, we have to administer the same test to the same source.

CHAPTER 14

FORM DESIGN AND LAYOUT

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

The front page has a title, instructions, and seal. The text page 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 perceive more than seven to nine words in a single glance. A string of seven to nine words with the type size we use usually takes up 3-1/2 inches. Furthermore, the two-column format gives the page a more formal and patterned look.

To reduce bulk, both sides of a page are 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 form may or may not have a cover.

INSTRUCTIONS

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

1. State the purpose of the survey.
2. Explain what GAO is, the basis of GAO's authority, and why GAO is conducting the survey.
3. Tell how and why respondents were selected.
4. Explain why their answers are important.
5. Tell how to complete the form.
6. Provide mail-back instructions.
7. List the person to call if help is needed to complete the form.
8. Provide assurances of confidentiality and anonymity.
9. Tell how long it will take to complete the form.

10. Explain how the data will be used.
11. Explain who will have access to the information.
12. Disclose uses that may affect respondents.
13. Present the response efforts as a favor and thank respondents for their cooperation.

The instructions should be concise, courteous, and businesslike.

FORM PREPARATION

About two thirds of GAO questionnaires and most pretests are reproduced from texts prepared on word processors. Although they are not usually as attractive and readable as texts prepared by commercial printers, they are quicker and cheaper to produce. For most questionnaires, a well-designed word-processor format is adequate. However, an attractive, readable, and businesslike style and type should be used. The formatting guidelines for the composition of technical text also apply to questionnaires prepared on word processors. Documents that look official, professional, and inviting are likely to be read. Good layout and composition can cut reading time in half and can reduce the respondent's burden. We use typesetting when

- the respondent group has low literacy,
- the questionnaire is very long and complex,
- we are surveying a large population,
- we are addressing a prestigious group, or
- our professional image is very important.

Type style and design must be specific and the questionnaire forwarded to the visual communications branch, which arranges for the composition and returns the proofs. This process takes 1 to 2 weeks and costs \$50 to \$100 per original page.

THE STYLE OF THE FORM

We use the size, style, and density of type as signposts to guide the reader's eye and to signal the kind of information being presented.

As shown in the samples on pages 137-40, the title is the most noticeable feature on the front page. It is a short statement (12 words or less) that should identify the population from which information is sought and give a clear idea of what the questionnaire is about. Because of its importance, we use large type, 14 point, in bold. (A point is 1/72 of an inch.) We use Universal or a similar typeface because it is official



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

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

INSTRUCTIONS

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 than 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.

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

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) _____

Italicized text style type

(1-5)
(6)

5pt. Gothic italics

REDUCED, NOT ACTUAL SIZE

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 provided during your workday. We also need time and cost information. Please list the usual number of hours of 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 hr) | | |
|--|---|-------------------------------------|------------------------|----------------------------|--|-----------------|--|---|---|
| | By spouse or relative at your home | By spouse or relative at their home | By sister at your home | By sister at sister's home | By staff at child care center, nursery school, kindergarten or group home* | Other (Specify) | Amount | | Total Cost Per Week (Include relative expenses) |
| | | | | | | | Total Number of Hours of Care During Workday | Total Number of Days Per Week Child Care Provided | |
| | 1 | 2 | 3 | 4 | 5 | 6 | | | |
| 1. (yrs) (mos) (25-27) | (28) | | | | | | (29-30) | (31) | \$ (32) |
| 2. (yrs) (mos) (35-37) | (38) | | | | | | (39-40) | (41) | \$ (42) |
| 3. (yrs) (mos) (45-47) | (48) | | | | | | (49-50) | (51) | \$ (52) |
| 4. (yrs) (mos) (55-57) | (58) | | | | | | (59-60) | (61) | \$ (62) |
| 5. (yrs) (mos) (65-67) | (68) | | | | | | (69-70) | (71) | \$ (72) |
| 6. (yrs) (mos) (6-8) | (9) | | | | | | (10-11) | (12) | \$ (13) |

REDUCED, NOT ACTUAL SIZE



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.

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.) | | | | | | (16) |
| 2. Hours and days service available or months of the year | | | | | | (17) |
| 3. Convenience of services (travel time and distance) | | | | | | (18) |
| 4. Safety and well-being of children | | | | | | (19) |
| 5. Quality of care, staff, program and facility, etc. | | | | | | (20) |
| 6. Cost of care | | | | | | (21) |

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) _____ (one-way trip time in minutes)

(22-24) (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

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
4. ☐ Of moderate interest to me
5. ☐ Of great interest to me
6. ☐ Of very great interest to me
- (GO TO
QUESTION 31)
- (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

looking and easy to read in bold capital letters. (Usually, capital letters are much more difficult to read than lowercase letters.)

The next feature the reader sees is GAO's 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 reader's eye next, are short phrases that tell what each part of the form is about. They will stand out if they are set in 12-point Universal bold and 11-point Universal demi-bold or similar typefaces.

Most of the form is text containing the instructions, questions, and answers. Here, we usually use 9-point or 10-point Times Roman, Baskerville, Press Roman, or similar type. These are clear, easy-to-read, official-looking typefaces, 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 readers 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 reader 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, because skips are very important, and substantially reduces 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. We prefer bold rather than underlining, because underlining stops the eye movement and slows the reader down.

Next comes the response space--little boxes to check; a column, row, 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 justified, or aligned, to the left of the response. Rows or columns or column-row matrixes are justified to the right, so that they line up with the row and column headings. 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, but sometimes the column headings are in Gothic, because it can be squeezed more than most other typefaces. It also reads well for very short passages.

All questions and response alternatives are numbered rather than lettered. These numbers double as codes for 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. They are not big enough to distract the respondent but not too small for the keypunch operator to read.

Shading is used to fill in space that the reader 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 form, and makes the form look more inviting. We try to give the reader as much white space as possible without expanding the number of pages.

CHAPTER 15

PREPARING THE MAIL-OUT PACKAGE AND COLLECTING AND REDUCING THE DATA

After the questionnaire has been developed, several tasks still have to be completed, as summarized below:

- Develop a computerized mailing list, a cover letter, and other mail-out materials and assemble the mail-out package.
- Monitor returns, conduct follow-ups, and make prekeypunching edits.
- Key punch the questionnaires and verify the computer file.

PREPARATION OF THE MAIL-OUT PACKAGE

Before mailing the questionnaire to potential respondents, we need to develop a computerized address file, prepare a cover letter, and assemble other materials (such as return envelopes) for the mail-out package.

Address files

While designing and testing the questionnaire, we were also selecting sample cases for inclusion in the survey. (See chapter 5.) Now, a computerized list of the addresses of all the sample units we selected must be prepared.

We normally begin with a hard-copy list of addresses, either GAO-constructed or obtained from another source. 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,
- job titles (as appropriate),
- titles (Dr., Ms., Mr.),
- 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 keypunched on tape and the tape 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 and the reviews of it can be repeated and corrections made. The corrected file should then be put into a special GAO program that assigns a unique case number to each sampled unit and puts the file in this format:

Mr. Thomas Fentworth
226 Whitehall Blvd.
Rochester NY 14617

At this point, a hard copy list should be printed for use in controlling mailed and returned questionnaires.

Cover letter

Almost as important as the questionnaire itself is the cover letter that accompanies it. Because respondents see the cover letter first, their decision to participate in the survey is often made on the letter's strength. Therefore, the letter should incorporate the following guidelines, which have been found to increase the likelihood of a reply. (A typical cover letter following for these guidelines is also shown.)

1. Design the mail-out package so that the cover letter is the first thing seen.
2. Have the letter neatly typed, not printed or xeroxed.
3. Use an official-looking format and style of writing but avoid being impersonal, ambiguous, or unclear.
4. Send the letter by first-class air mail, when appropriate. (The return envelope should also be for first class.)
5. Address each individual in person.
6. Explain what GAO is and its legitimate role in collecting this data.
7. Without being pretentious, imply that GAO is an important agency with influence.
8. State the purpose of the project.
9. Stress the importance of the project.
10. Relate the project to the respondent.
11. Emphasize the importance of the respondent or the respondent's organization.



UNITED STATES GENERAL ACCOUNTING OFFICE

WASHINGTON, D.C. 20548

PROGRAM EVALUATION
AND
METHODOLOGY DIVISION

January 10, 1985

Mr. Ronald Jones
St. Boulevard Road
Cleveland, Ohio 20698

Dear Mr. Jones:

The U.S. General Accounting Office--the agency of the Congress charged with the investigating the use of federal funds--is currently reviewing the effectiveness of your National Guard or Reserve training. Our review, along with others we have made in the past, is aimed at improving the nation's overall military readiness. We could not undertake this review without first considering the opinions and experiences of the people like you who staff this effort.

Of course, we would have liked to talk to each of you in person, but as you may realize, this is impossible. Therefore, we have selected a sample of people who, like yourself, represent a cross-section of the forces and are asking them to complete a short questionnaire. Although this questionnaire should take only 15 or 20 minutes to complete, your answers are of vital importance to our review and to others like yourself who are currently serving their country in this program. In case you are wondering "why me?" your name was chosen at random as part of a sample.

Since the sample represents a very small portion of service personnel, we must hear from everyone we have asked for help or our results will not represent a true cross-section of service men and women.

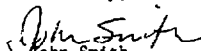
We need your frank and honest answers and we want to make one point clear. Your answers are confidential and will be used only for the purposes of this study. They will not become part of your service record or any other file. In fact, your name and address will be disassociated from your questionnaire and your participation will not be known. Nobody will be able to tell how you or any other person answered. Remember, while your name is not important to the results of this study, your experiences and opinions are. We cannot make meaningful recommendations without help and consultation from you and others like you.

It is essential that you complete the questionnaire and return it in the enclosed envelope within 10 days, if possible.

If you have any problems with this questionnaire, please call Brian Keenan at (202) 275-3762.

Thank you for your cooperation.

Sincerely,


John Smith
Regional Director

Enclosures

12. Stress the importance of the answers and the study to the respondent and the nation. If possible, make references to possible benefits to respondents.
13. Tell how and why the respondent was selected.
14. State that the questionnaire can be answered easily and in a short time. Tell truthfully how long it will take to complete.
15. Emphasize the importance of replies from everyone in the sample and express your dependence on those replies.
16. Ask a favor.
17. Guarantee the respondent anonymity or confidentiality and state that responses will have no hidden uses.
18. Ask for honest and frank answers.
19. Urge prompt responses.
20. Warn about a follow-up for those who do not reply.
21. Mention the possibility of a verifying personal interview when appropriate.
22. Provide a name and a phone number in case the respondent needs assistance in completing the form.
23. Express your appreciation.
24. Have the letter signed by hand in blue ink by the highest appropriate echelon of responsibility. If many letters are to be sent out, have several clerks sign them.

One item--our pledge of confidentiality--is worthy of further discussion. Some GAO studies are enhanced (by higher response rates and more honest answers) by telling potential respondents that their responses will not be reported in a form other than as part of aggregate statistics. But before a pledge of confidentiality is used, a written justification should be prepared and approved by the division director. For work being done for the Congress, our pledge should be approved in writing by the requestor. Current policy guidance on pledges of confidentiality is included in GAO's general policy manual, chapter 7, appendix III.

Once the cover letter has been written, edited, reviewed, and revised, it is ready to be typed into the computer system as a separate file. GAO has a computer program to produce the

cover letters by merging each address in the address file with the cover letter file. At this point, the letters are ready for signatures.

Other mail-out materials

Before the cover letters are run, 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 are used only when the respondents remain anonymous (that is, when the questionnaires contain no identifying codes). The cards, which are returned separately from the questionnaires, tell us that the respondents have completed the questionnaires without associating them with their questionnaires.
3. Envelopes for the mail-out package usually have windows.
4. Occasionally, letters of endorsement from influential people are included in the mail-out package when we believe 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 are done by an outside firm when we have a long lead time and a large sample and when the benefits outweigh the costs.

We usually distribute the packages directly rather than 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 VA hospital administrators to distribute to the physician staff, and in another we gave them to union leaders to give to their members. Both surveys had to be discounted because of poor response rates and uncontrolled sample selection.

DATA COLLECTION

Essential to a good data collection phase is the monitoring of responses and nonresponses and a continuing effort to get the

responses. The response rate goal for GAO surveys is usually 75 to 90 percent.

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.

The earliest returns may be undeliverable packages. For each undeliverable, we should note on the control list why the package could not be delivered. Incorrect addresses should be recorded and new mailings prepared when feasible. Other early returns may come from those who were erroneously included in the sample and therefore do not complete the questionnaire. It is important to separate inappropriately sampled units so that we can adjust both the sample size and the universe size. 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 be followed by one or two follow-ups, using the normal mail 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 will be likely to trail off at about a rate of 30 percent to 50 percent. At this point, 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, about 3 weeks after mailing the first follow-up, we should 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. We may be able to stop, or we may want to try one last follow-up by mailgram or the 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 (mailgrams can be costly).

Follow-up letters are prepared and produced in a manner similar to the preparation of the original cover letter (shown in a sample here). The mailing list is adjusted by eliminating from



UNITED STATES GENERAL ACCOUNTING OFFICE

WASHINGTON, D.C. 20548

PROGRAM EVALUATION
AND
METHODOLOGY DIVISION

January 10, 1983

Dr. John Doe
1776 Main Street
Middletown, Pennsylvania 11234

Dear Dr. Doe:

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 return to complete our review. Your opinions regarding ESRD physician 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. Your responses will be combined with those of other physicians for our report to the Congress. The answers of individual physicians will not be identified.

We have enclosed another copy of our questionnaire for your convenience. If you have any questions, please call Bob Sayers or Maureen Driscoll collect at (303) 964-0052.

Thank you for your cooperation.

Sincerely,

Louis Lucas
Acting Manager
Boston Regional Office

Enclosure

it the names of those who responded, 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.

Prekeypunching edits of responses

As questionnaires are returned, they must be edited before they can be keypunched and entered into the computer system as a file. This editing process can take weeks to complete, but the project team can begin editing as soon as responses are received. Editing should not 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 the keypuncher?
4. Do the open-ended responses provide useful data?

Inadequate responses must be eliminated, corrected according to the evaluators' judgment, 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 we can analyze the data, we must move it from its current hard-copy form (the questionnaire) into a computerized data file that accurately reflects the hard-copy data. We begin with keypunching.

Keypunching

Keypunching for GAO questionnaires is normally done by an outside contractor. Nearly always, the keypuncher punches from one of two sources--a coding sheet laid out in an 80-column card image format and prepared by the project team or the questionnaires themselves. Many GAO questionnaires are recoded (80-column card format) for keypunching. The keying is generally done onto a tape that can readily be entered into the computer system as an unedited raw data file. Keying instructions unique

to the individual job are provided to the keypunchers for guidance. Two of the project team's primary tasks are to ensure that the questionnaires given to the keyers are punched and that all original questionnaires are returned--a control function.

Loading the unedited raw data file

In this short but necessary step, the tape containing the unedited raw data file is loaded in the computer system. This is normally done overnight with the aid of a system operator.

Keypunch verification

Once loaded in the computer system, the unedited raw data file can be converted to hard copy, in order to verify 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 punched again.

Rather than verify the entire file, we can take a sample. How large should the sample be? Large enough to statistically ensure, at the 95-percent confidence level, that the error rate is not more than 1 percent. This usually amounts to a 10-percent sample of cases or 40 cases, whichever is less. The verification process works best when two evaluators work together; one reads from the questionnaire while the other views the printed computer file.

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 a 1 or a 2 in the file, 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 completing this process, we have an edited raw data file that can be used in the initial steps of the analysis phase, as discussed in the next chapter.

CHAPTER 16

ANALYSIS OF QUESTIONNAIRE RESULTS

Although the focus of this paper is on data collection, we need to make a connection to the next step--data analysis. Actually, the concern with data analysis occurs very early in the project planning process. When we are thinking about the overall evaluation questions and begin to develop the design for answering the questions, we will logically be led to the point of considering data analysis.

For example, if the evaluation question we are considering is descriptive, we may decide that a simple descriptive statistic like a mean and its corresponding confidence interval may suffice. However, the point of this chapter is not to get into the details of data analysis but, rather, to sketch out several main modes of analysis. Much more information will be found in a forthcoming PEMD transfer paper to be entitled "Introduction to Data Analysis."

ANALYSIS PLAN

A data analysis plan should be developed as part of the evaluation design and long before any data are collected. Planning forces us to decide what kind of findings we do and do not need to complete our evaluation. This rendering process is important, because it is very easy to overburden the study with unnecessary analyses.

On a typical project, most standard analysis packages can provide millions of analyses that would take many years to interpret. We have to run the analysis; otherwise, it will run us. Also, unplanned analysis can result in fishing or data dredging, when evaluators run analyses without regard to a design or preconceived reason, just to see what they will get. This is not science but chance, and such methods have little credibility.

Thinking through the data analysis may cause us to reconsider our data collection plan or even the evaluation questions themselves. In planning the data analysis, we might realize, for example, that we need an additional piece of data that we had not thought of before. The selection of analysis techniques and the variables to be analyzed will, of course, be determined by the evaluation questions and the design requirements. We also need to make sure that our statistical analysis software routines can satisfy these requirements. For example, can they handle the size and number of variables? And can they do the analyses required?

Later, when the analysis begins, we will know how adequate our planning and data collection have been. If the measures were properly defined, relevant, and sound, and if the data relationships turn out as expected, then the analysis will

proceed as planned. Usually, however, projects are imperfect and there are some gaps in the planning and flaws 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. We need then to modify our analysis plan, scaling back the effort, expanding it to cope with unexpected developments, or exploring different ways of answering the evaluation questions. Regardless of departures from our original plan, however, the analysis must still proceed in a logical, step-by-step fashion from very simple analyses to a limited number of more complex analyses.

ITEM RESPONSES AND UNIVARIATE ANALYSIS

The first step is to go just a short way beyond the raw data on questionnaires by producing what is often called a "code book." The code book tells us 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, we can compute descriptive statistics such as rank order scores, measures of central tendency (averages, modes, and medians), deviations from the central tendency, and other indicators that help describe the frequency distributions.

BIVARIATE ANALYSIS AND COMPARISON OF TWO GROUPS

Here, we begin to look at the relationship between two variables or make comparisons between groups of respondents. If we want to study the relationship between two variables, we use correlational techniques. These techniques show that a change in one variable is associated with a change in another. For example, we might want to determine whether the performance of the Federal Aviation Administrator'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 crosstabulations, chi-square comparisons, t tests, and analyses of variance. These and other analytic techniques will be examined in a forthcoming PEMD transfer paper to be entitled "Introduction to Data Analysis."

MULTIVARIATE ANALYSIS AND COMPARISON OF MULTIPLE GROUPS

We use this level of analysis when we want to 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 FAA 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 or discriminant analysis.

CHOICE OF ANALYSIS METHODS

The choice of data analysis methods depends largely on the evaluation questions and subject matter under study. For example, if we had a question about whether the performance of FAA 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 we study 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 we have a choice between using associations and using group differences, and sometimes we 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.

BIBLIOGRAPHY

- 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.
- Bradburn, M. N., and S. Sudman. Response Effects in Surveys. Chicago: Aldine, 1974.
- 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.
- Erdoes, P. L. Professional Mail Surveys. New York: McGraw-Hill, 1970.
- Flesch, R. Say What You Mean. New York: John Wiley and Sons, 1974.
- Lockhart, D. C. (ed.). Making Effective Use of Mailed Questionnaires. San Francisco: Jossey-Bass, 1984.
- 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.
- Schuman, H., and S. Presser. Questions and Answers in Attitude Surveys. New York: Harcourt Brace Jovanovich, 1981.
- Sudman S. Applied Sampling. New York: Academic Press, 1976.
- , and M. N. Bradburn. Asking Questions. San Francisco: Jossey-Bass, 1982.
- Warwick, D. P., and A. C. Lininger. The Sample Survey: Theory and Practice. New York: McGraw-Hill, 1975.

GLOSSARY

Bias. The extent to which estimates or measures systematically underestimate or overestimate a true value.

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 a characteristic or attribute or variable relationship. The concepts are often unobservable ideas or abstractions such as community context or performance.

Estimation error. The amount by which an estimate differs from a true value. This error includes the error from all sources (e.g., sampling error, measurement error, and so on).

Judgment sample. A sample selected by using discretionary criteria rather than criteria based on the laws of probability.

Measurement. An observation procedure for assigning a value to a variable.

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 either answer 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. It is different from verification and validation.

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 universe 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.

Validation. The procedures necessary to demonstrate that a question or questions are measuring the concepts that they were designed to measure.

Variable. A property, characteristic, or attribute that can be measured and can vary from one case to another.

Verification. An effort to test the accuracy and soundness of the measurement data. It can be different from validation in that the concern is with accuracy rather than with the proof of the measurement concept.