CHAPTER 2

RESEARCH DESIGN AND DATA COLLECTION

BACKWARD RESEARCH: IN MY END IS MY BEGINNING

We will turn shortly to the strengths and weaknesses of different research designs and data collection strategies that give us the raw materials that our statistical tools will craft into effective and honest arguments and presentations. But we will begin by considering where we want to end by using Alan Andreasen's (2002) principle of backward research.

Andreasen (2002) tells a story that shows us what can occur when an analyst works with a client in the public or nonprofit sectors. His example comes from a nonprofit performing arts organization, but it surely applies far beyond this case. In short, the executive director of the organization wants to know more about her organization's market and commissions a research agency to collect, analyze, and report on the several types of audience that attend performing arts programs in her city. The agency consults previous studies of this kind and then designs, say, a random digit dial (RDD) telephone survey of adult residents of the city, conducts interviews, creates data files, and applies the appropriate statistical tools to analyze the data. The agency then produces a glossy report, using the finest graphical tools available. It's filled with statistics.

The executive director reads the report and, having finished it, lays it aside with a sigh of disappointment. The report tells her what hundreds of similar studies have already reported: The audiences for the performing arts are largely female, well educated, urban, relatively young or old, and well-off economically. She already knew that.

Even the few new nuggets of knowledge in the report did not provide her with insights on which she could act. Why, for example, do men attend less often than women? Why are the middle-aged less likely to attend? How could she motivate these groups to come more frequently while sustaining her loyal market segments? How could she create and sustain loyal audiences in the future? The executive director, in Andreasen's telling of the story, explains in frustration, "When the researcher tried to explain the results, it was obvious he hadn't understood what I wanted. The results were all a bit off the mark" (pp. 60–61).

Clearly, part of the fault lies with the executive director in not communicating what she really wanted to know. As explained in the previous chapter, the most successful research is marked by a close collaboration between researcher and client, where the research is driven by the client's questions and needs (in contrast, e.g., to the questions that a scientific discipline strives to answer).

The story above is commonplace. A group of my students had volunteered to help the Graduate Student Association (GSA) draft a questionnaire to be administered to graduate students in professional degree programs in the School of Arts and Sciences at the University of Pennsylvania. I reviewed the draft questionnaire and asked the students, rather skeptically, what decisions the GSA planned to make on the basis of the answers to the draft survey's large number of questions about levels of satisfaction with issues ranging from advisors to debt. I couldn't see how they could put the data to any use. "Well," the response from one of the students came, "GSA surveyed the PhD students about a year ago, and they thought they should survey the professional degree graduate students too. GSA's president told us he'd figure out how to use the data after the results had come in." Ugh.

It is, of course, the case that a client does not always know exactly what he or she needs to know. But mindless fishing expeditions are unlikely ever to catch a fish worth eating. Be more "planful" and disciplined and help your client be so too by following a "backward research design."

Here are the 10 steps of a backward research design (modified from Andreasen) that will ensure more successful outcomes than the ones described above:

1. Ask your client what key decisions are to be made using the research's results.

2. Determine—in collaboration with your client—what information will help him make those decisions.

3. Prepare a prototype report or ask your client what actions or decisions would follow if you discovered x, y, or z. This exercise will also

help you determine what alternative explanations need to be considered (and perhaps discarded).

4. Determine what questions must be answered to complete the final report.

5. Ascertain whether these questions have already been answered in other research.

6. If not, design one or more studies that can be conducted within your time and financial constraints.

7. Implement the study(ies).

8. Write the report and present the results.

9. Help deflect or respond to any criticism the report might receive and direct any praise toward your client.

10. Evaluate the research process and contribution, and propose a new study to answer the new questions your client now comes up with but couldn't have foreseen before your brilliant study and presentation of results.

Interestingly, the first five of these steps do not require you to raise a statistical finger (although you may be required to know how to read and evaluate someone else's research, a skill that follows from doing research yourself).

Ideally, you also have a theory or model, a story that offers an explanation of the patterns of relationships you expect to find in answer to questions such as why middle-aged men tend not to attend the ballet, why homicide rates in the United States dropped in the 1990s, why poor children perform more poorly on standardized tests than wealthy ones, or why some women on welfare tend to stay on welfare for extended periods of time. Moreover, your theory and the questions you seek to answer—should you need to collect your own data—will build on research that has already established the plausibility of at least parts of your theory or model. If so, you're humbly following in the footsteps of Sir Isaac Newton, who, in a letter to a rival mathematician, Robert Hooke, wrote in 1676, "If I have seen further it is by standing on the shoulders of giants." (Newton may not have been beyond sarcasm here, as Hooke himself was rather short of stature.)

It is, of course, possible that extant research does not provide a solid foundation for what you seek to answer. You may find yourself in the situation of the two statisticians who were having a hard time solving a particular problem. "You know what our problem is Bill [Kruskal]?" asked Fred Mosteller. "We're standing on the shoulders of pygmies."

Hopefully, you've already picked up some skills about how to find existing research on the subject and questions of interest to you. Technologies such as the Internet and Google have made these tasks easier, and libraries at major research universities provide valuable resources and skilled reference librarians to help you in your search. These resources make the search and acquisition of statistics and data (observations on which statistics are based) relatively easy (although you may be surprised to discover that despite the vast amount of information and data that exist today, answers to your specific questions may not reside among these data). We will practice, in the exercise to be found at the conclusion of Chapter 3 of this book, how to bring data collected by someone else into Excel or SPSS for further analysis. There's plenty of data out there for your examination. A list of some of these possible sources can be found in Appendix A, "From Whence Do Data Come?"

Let's assume that you've made it through Andreasen's first five steps and have concluded that the available studies—although providing a portrait of the conditions in which you are interested and pointing toward a variety of causes and responses you may want to investigate—don't give you or your client the answers needed to make more informed decisions. You have to design a study or prepare a Request for Proposal (RFP) to the research community in order to conduct a study that will answer the questions you've posed. (Note here too that although someone else may actually conduct the study, complete the analysis, and submit report(s), you and your colleagues may be required to evaluate the credibility of the proposals for this research before awarding a contract and then assess the quality of the data and reports that others eventually produce.)

Not surprisingly, it turns out that different research designs have different strengths and weaknesses that you'll have to consider in conducting the study yourself or in outsourcing this task. This is the topic to which we now turn.

● STRENGTHS AND WEAKNESSES OF DIFFERENT RESEARCH DESIGNS

What are the basic types of research designs that you might turn to in order to answer the questions that will assist you or others in making more informed choices? They are as follows:

- Experiments
- Sample surveys
  - Mail, phone, face-to-face, Internet, computer assisted
  - Longitudinal, cross-sectional

- Administrative records
- "Actors," "confederates," "testers," and "audits"
- Observations (unobtrusive or participant)
- Focus groups

Whether any of these research designs draw a sample from a population of interest is a somewhat different, and central, concern of **inferential statistics**, a topic to which we will turn in Chapter 8. Suffice it here to note that there are occasions when it is possible to study an entire population in contrast to a sample drawn from it (e.g., all students in a school, all school districts in a state).

Beware as well that you are likely to draw the impression from the following pages that policy research is an impossible task, fraught with problems that produce errors at every stage. While it is true that no data collection strategy or design is perfect, don't allow these difficulties to lead you to despair of ever finding answers to questions that will help you or someone else make better decisions. Policy research actually does help (and, of course, sometimes it doesn't). There are better and worse ways to gather data. There are some data and statistical tools that will give you more or less confidence in your conclusions, despite knowing that your data harbor imperfections. And there are ways in which you can combine different research designs within the same study (e.g., embed an experiment within a cross-sectional survey), which helps mitigate the weaknesses of any one research design (Boruch, 1975).

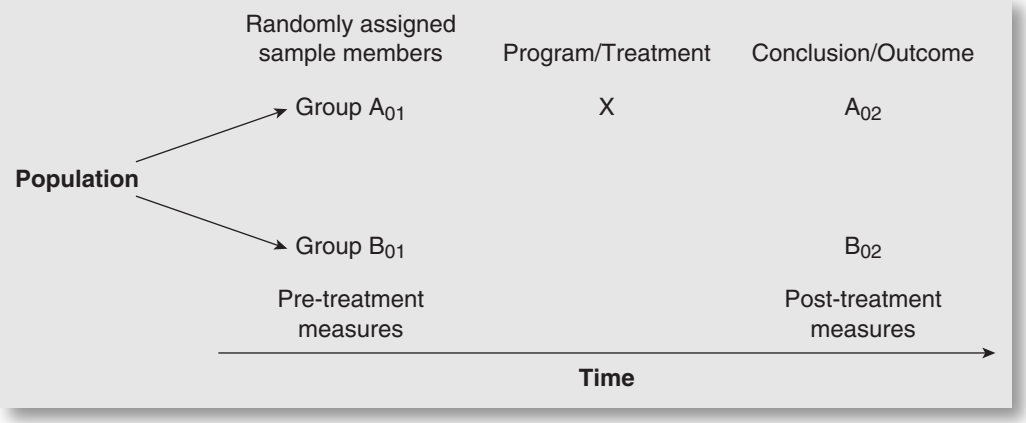
Experiments

Some research and policy wonks admire experiments so much as to label them the "gold standard" of research. Those who champion the virtues of experiments are somewhat correct. But we can't always use gold or afford it, as we will note in our discussion of the weaknesses of experimental designs.

Figure 2.1 presents a diagram of a classical experimental design.

We begin with a population of interest to us. Let's say it's the noninstitutionalized adults in the United States. (*Noninstitutionalized* is a term used by survey researchers typically for the purpose of excluding from a study people who are difficult to reach and interview. These may include, for example, patients in hospitals, prisoners, and military personnel stationed abroad.) Let's further stipulate that we want to test whether the health insurance provided by the federal government results in greater

Figure 2.1 A Classical Experimental Design



access, better quality, and lower costs than the existing, employer-based, private health insurance plans. This is an important question for which no experimental data exist, although at least one social experiment in health care—the RAND health insurance experiment, which was conducted between 1974 and 1982—used randomized assignments to assess the consequences of different levels of copayment in private health insurance plans on the use of health care services. This research demonstrated, not surprisingly, that the more people had to spend directly out of their own pocket, the less health care service they consumed. What you may not have considered is that the experiment also showed that health among the sick poor was adversely affected with higher copays (Relman, 2007, pp. 101–103).

You've probably already identified some difficulties with assigning some randomly selected, noninstitutionalized adults for enrollment in a national health insurance program in our hypothetical example. You might, for example, decide to screen from your initial population those people currently receiving some form of national health insurance. Estimates range as high as 45% of the American public currently receiving such health insurance through Medicare, Medicaid, State Children's Health Insurance Program (SCHIP), or military and veterans' hospitals. Among other implementation problems, such a study would require nearly half of your subjects to drop their current private health care insurance. The currently uninsured—about 17% of the American population—might very well welcome participation in the experiment. (The RAND study provided "side payments" to participants in the experiment who suffered some financial loss as a consequence of their random assignment to different health insurance plans.) Our own hypothetical experiment would ask that the sample of respondents randomly assigned to a government health

plan faithfully stay in the program—which statisticians refer to as the "treatment." In the diagram, Group A is a sample randomly drawn from the population that receives the treatment. (We will discuss random selection and sampling later, in Chapter 7.) Another group of people, Group B, are also randomly selected but receive nothing different from their current coverage (private or none). This is often called the "control" group.

We gather information about current health conditions from everyone who was randomly selected for the study, before the assignment to the treatment and control groups. If the assignment was properly executed, we can be confident that the two groups won't differ on these measures at their initial assignment to the treatment or control group (but it's good to check anyway). If we discover that they don't differ, we're also even more confident that the randomization has produced two groups that do not differ on things we did not measure. We administer the treatment and then, after an adequate amount of time elapses for effects to appear, remeasure the outcomes of interest. If Group A differs from Group B at the conclusion of the study (e.g., fewer emergency room visits, fewer deaths), we would conclude that it was the treatment that caused the difference. National health insurance is less expensive, more broadly accessible, and provides better care! Obviously, we should pass such national legislation. If only life were so simple!

As we'll also see in survey samples, random selection is important. In the case of experiments, it permits us to isolate the unique effects of the treatment from all other possible factors that influence health conditions. (You can test more than one treatment or program at a time using factorial designs, but these are beyond the scope of this book. For an intuitive understanding of such designs, see Almquist & Wyner, 2001.) If the randomization is properly executed, the two groups will differ only in that one received the treatment and the other did not. Any difference we observe in outcomes is attributable to the treatment because it is the only condition on which they differ.

Obviously, experimental designs are commonplace in areas such as the assessment of pharmaceutical products and medical devices. More than 1,000,000 randomized control trials (RCTs)—another term for "experimental design"—of medical treatments have been conducted since the late 1940s, when the first antibiotic treatment for tuberculosis, streptomycin, was evaluated through an experimental design (Sherman, 2003, p. 10). Many of these studies have been identified, cataloged, and themselves studied through the Cochrane Collection (www.cochrane.org/index.htm). Although much less frequent in the social sciences, more than 10,000 experiments are known to have been conducted and are referenced in the Campbell Collaboration (www.campbellcollaboration.org/). Experiments are especially useful in determining "what works."

Experiments face a variety of challenges, however. Opponents of experimental designs may argue that it is ethically unjustifiable to provide some benefits to one set of people while denying them to another. This objection does not always hold water, however. Surely, if we knew the precise effect and costs and benefits of a program, we wouldn't need to conduct an experiment. But rarely do we know this. Social scientists have also discovered that some programs that are believed to result in beneficial outcomes actually do not (e.g., many self-help programs) and, worse, have been shown in randomized controlled trials to do harm to the treatment group.

For example, RCTs on the program D.A.R.E. (Drug Abuse and Resistance Education) have shown the program to be ineffective (U.S. General Accounting Office, 2003), although a new D.A.R.E. curriculum for seventh and ninth graders, "Take Charge of Your Life," is currently undergoing a 5-year randomized trial evaluation by the University of Akron. Some studies have shown that kids exposed to the earlier D.A.R.E. program ended up taking more illicit drugs than the control group of randomly assigned youth who did not participate in the program. These "boomerang" effects, as they are sometimes called, arise when subjects interpret and/or evaluate the likelihood of an outcome differently than what the evaluation or program designer believes to be the case (Capella, Yzer, & Fishbein, 2003).

More generally, ethical/legal objections to social experiments can be mitigated through any one or more of the following strategies (Boruch, 2004, p. 5098):

- Include in the study only those groups of people for whom the treatment's effectiveness is uncertain or not legally or ethically questionable.
- Employ a wait-list in which those currently denied the treatment are given it later (assuming that it has been shown to be effective or, at minimum, not harmful after the initial results of the first set of assignments are known).
- Assign entire institutions (e.g., schools, hospitals) or geopolitical units (e.g., police precincts, counties) to treatment and control groups rather than assigning individuals within those units (see Boruch, 2005).

Moreover, if the treatment is in short supply relative to its demand, random assignment to a treatment can be seen as a fair means of allocating the treatment.

There are, of course, a large number of important questions that simply can't be assigned to an RCT. Questions in international relations do not lend themselves to random assignments of countries or their leaders to control and treatment groups. One cannot, for example, assign governments to be either

democratic or authoritarian and then observe whether the democratic countries are less likely to engage in war than the authoritarian ones, a widespread hypothesis in international relations theory (Frieden & Lake, 2005).

Further questions arise concerning the "fidelity" of programs modeled on an RCT that have shown a treatment to produce a substantial and intended effect. An RCT of the effects of the mentoring provided by Big Brothers Big Sisters in a study by Public/Private Ventures (Tierney & Grossman, 1995), for example, demonstrated the positive effects of adult mentoring on the kids whom the program served. Another RCT of the long-term effects of early child care intervention—the Ypsilanti Perry Preschool experiment—found higher rates of high school graduation and employment and lower rates of teen pregnancy and arrest (Weber, Foster, & Weikart, 1978). Similar results to these were found in an RCT of black children in Harlem (Deutsch, 1967).

Other programs have been subsequently based ostensibly on these successful programs and cite one of these experiments to justify their own. This has been the case even when these subsequent programs lack key elements of the studied programs (e.g., they don't screen and train "Bigs" as rigorously as Big Brother Big Sisters or don't provide the range and depth of support services to kids and their families of the Ypsilanti study).

"Gold standard" research designs can also be perverted, either in their implementation or in their interpretation. This is especially likely to be the case when large amounts of money are at risk. The RCT that reported Vioxx to be a safe and effective treatment for pain, for example, demonstrated two fundamental flaws in its clinical trial. The first involved fraud. The second used the smoke screen of "statistical significance" to deflect attention from its real, and troubling, effects. The fraudulent action was the omission of three people from the study who took Vioxx and suffered heart attacks. The report said that only five of the treatment group suffered heart attacks in contrast to only one participant from the control group. The second flaw was more subtle. The comparison of five to one heart attacks was reported not to have achieved the status of "statistical significance," a term to which we will return in Chapter 8 and the remainder of the book. Suffice it to say that the differences between five (or eight) heart attacks and one could well be substantively meaningful if not "statistically significant" (Ziliak & McCloskey, 2008, pp 28–31).

Most of the examples noted above employ what we might consider a classical experimental design.[1] But departures from the design features of classical experiments are commonplace and predate RCTs. Several authors have sought to order and classify these departures from experimental designs under the rubric of "quasi-experiments" (Campbell & Stanley, 1963; Cook & Campbell, 1979), although one of these authors has subsequently admitted to have come to call many such designs "queezy-experiments" (Cook, 2003).

Such designs include the following (Shadish, 2004):

- Interrupted time-series designs, in which consecutive observations are compared before and after the introduction of a treatment, which can be a program intervention or a change in policy (we will turn in more detail to a variety of such designs in Chapter 13)

- The outcomes of two or more treatments or conditions are studied, but an investigator does not control the assignment to these conditions. An investigator, for example, may sample from a list of people known to have participated in a job training program and compare their characteristics with those of a sample drawn from people known not to have participated in such a program. Such a design would be analogous to comparing A_{02} with B_{02} in Figure 2.1, without the initial random assignment to treatment and control groups. The results from such designs can be strongly influenced by selection bias; that is, people who volunteer to participate in a job training program may have unmeasured characteristics (e.g., ambition) that could produce any observed differences (e.g., employment) in outcomes rather than the treatment itself. Some studies of this kind can also produce results that are an artifact of program administrators permitting only those people most likely to succeed to participate in the program (a process referred to as "creaming"). Such processes are inherent in performance evaluation systems that look only at outcomes (e.g., the number of participants who get jobs within a month of graduation from a training program, without regard to their characteristics at program intake). This type of **quasi-experimental design** is often called a nonequivalent control group design.

- Single-case designs, in which one or more participants' responses to different dose levels of a treatment are observed over time

- Case-control designs, in which a group with a particular outcome or condition (e.g., lung cancer) among a set of "cases" is compared retrospectively with those without this condition who are otherwise similar (the "controls")

- Similarly, a matched comparison research design is constructed by, say, first collecting information from and/or about a group of interest (e.g., youth in a particular neighborhood) and then finding a sample of individuals from the broader population whose characteristics match the characteristics of the members of the first group. Joan McCord (2003), for example, conducted a 30-year follow-up study of research originally launched in 1942 among a group of boys who lived in two congested urban neighborhoods in Cambridge and Somerville, Massachusetts. Information about the group of boys from these neighborhoods was collected from the

boys themselves, their families, and the neighborhoods. A second group of boys was selected to match each member of the Cambridge-Somerville group on social background, temperament, and physique. A coin toss determined which boy from these matched pairs would receive weekly, and sometimes extensive, visits from caseworkers over a 3-year period. Thus, assignment to treatment and controls was random; selection for the sample itself was not. Selection for the second group depended on matching the characteristics of the members of the first group. (We will later see that this type of selection for a sample will require the use of a different set of tools to measure the strength and direction of a relationship from the tools used to analyze a survey where the participants are selected independently of each other.)

- Correlational designs, in which possible treatments, conditions, and outcomes are measured simultaneously without random assignment to treatment and control groups (this type of study often characterizes sample surveys, the design of which we will turn to shortly in more detail)

The point of this discussion on experimental designs, as well as the other research designs, is not that they can or cannot be trusted to produce sound and useful conclusions. The conclusions instead are that

- different designs have different strengths and weaknesses;
- the implementation of any design and the fidelity of its replication in later studies are the type of "devil in the detail" that can profoundly affect the results;
- one should take care in assuming that the results based on a sample from one kind of population (e.g., registered voters in Milwaukee) are applicable to a different population (e.g., registered voters in Philadelphia); and
- when the stakes are high, the incentives for manipulating the implementation of a study and distorting its interpretation are also high and require careful attention.

Sample Surveys

Surveys encompass a surprisingly wide variety of types. They vary, for example, by the methods for collecting responses: mail, phone, face-to-face, paper and pencil, or the Internet. They differ in that some capture responses on paper, others on the hard drive of a laptop. They vary by whether they collect information from a set of people at one point in time (often referred

to as a "cross-sectional" survey) or whether they collect information from the same subjects repeatedly over time (i.e., "longitudinal" or "panel" studies). Surveys can also collect data from and about organizations and events.

Sample surveys also vary by the ways in which respondents are selected. Selecting respondents randomly and with known probabilities of selection enables you to employ the vast array of statistical tools that we'll explore later in making inferences about a population from data collected only from a subset of that population (i.e., a random sample). In contrast, convenience samples (e.g., asking people to complete a survey in a shopping mall) or snowball samples (e.g., asking your current respondent to identify the next person to interview) provide no known probability of selection. Snowball samples may, however, be useful in studies of migration chains and social networks (see, e.g., Massey, Durand, & Malone, 2002). Although you cannot draw inferences about the characteristics of the population from which such samples are drawn, they may be valuable in generating ideas or pretesting a questionnaire. Such is also the case with the other types of research designs described below.

Sample surveys are used to collect information about an incredible variety of topics, although we often don't realize that the information reported to us on a nearly daily basis comes from this form of data collection. Unemployment rates (Current Population Survey), job counts (Current Employment Statistics), inflation (**Consumer Price Index, CPI**), consumer confidence (Survey of Consumers), illicit drug usage (National Survey on Drug Use and Health), academic performance of 4th, 8th, and 12th graders (National Assessment of Education Progress), risk factors for chronic diseases (Behavioral Risk Factor Surveillance System), and crime victimization (National Crime Victimization Survey) are just a few social and economic indicators that surveys produce.

There are also many surveys conducted for specific purposes—say, to understand how the public in a county views its local government and whether these attitudes vary by demographic group, geographic area, or the number of encounters citizens have with government workers and officials. You will see such a survey in the Orange County Survey of Public Perceptions as part of the exercises that accompany the chapters of this book.

If you're responsible for issuing an RFP to organizations whose purpose is to collect survey data (e.g., the National Opinion Research Center [NORC] at the University of Chicago, the Survey Research Center [SRC] at the University of Michigan, or a private firm such as Westat or Mathematica),[2] you will want to know (and ask the bidders for this work) the answers to at least the following questions about the survey being commissioned (see Groves et al., 2004, p. 33):

1. *How will the potential sample members be identified and selected?* (The identification of prospective sample members is referred to as the sampling frame—that is, a list of all members or an identifiable subset of the population in which you're interested.) The selection can be a simple random sample (SRS) from a list, say, of everyone in the phone directory of a town or city. The SRS is the most easily understood method of selecting sample members. But one often doesn't have a handy list of all members of a population. We'll return in Chapter 8 to a fuller discussion of selection techniques, including stratified and clustered selection methods that get around the absence of a list of all members of a population.

2. *What approach will be taken to contact those sampled, and how much effort will be devoted to trying to collect data from those who are hard to reach or reluctant to respond?* That is to say, how many times will the data-collecting agency return to a house, call a phone number again, or send another mail questionnaire or reminder before chalking the case up to a nonresponse? Will especially persuasive interviewers be assigned to those who are selected but who are reluctant to respond? Will incentives be offered to increase the participation of reluctant respondents?

3. *How much effort will be devoted to evaluating and testing the questions that are asked?* We will return in the next chapter to the issue of pretesting, to which this question refers. To how many people will a draft questionnaire be administered during the pretest? How will the respondents' comprehension of the question be assessed? How will we determine whether the questions are actually measuring what we believe (hope?) them to be measuring?

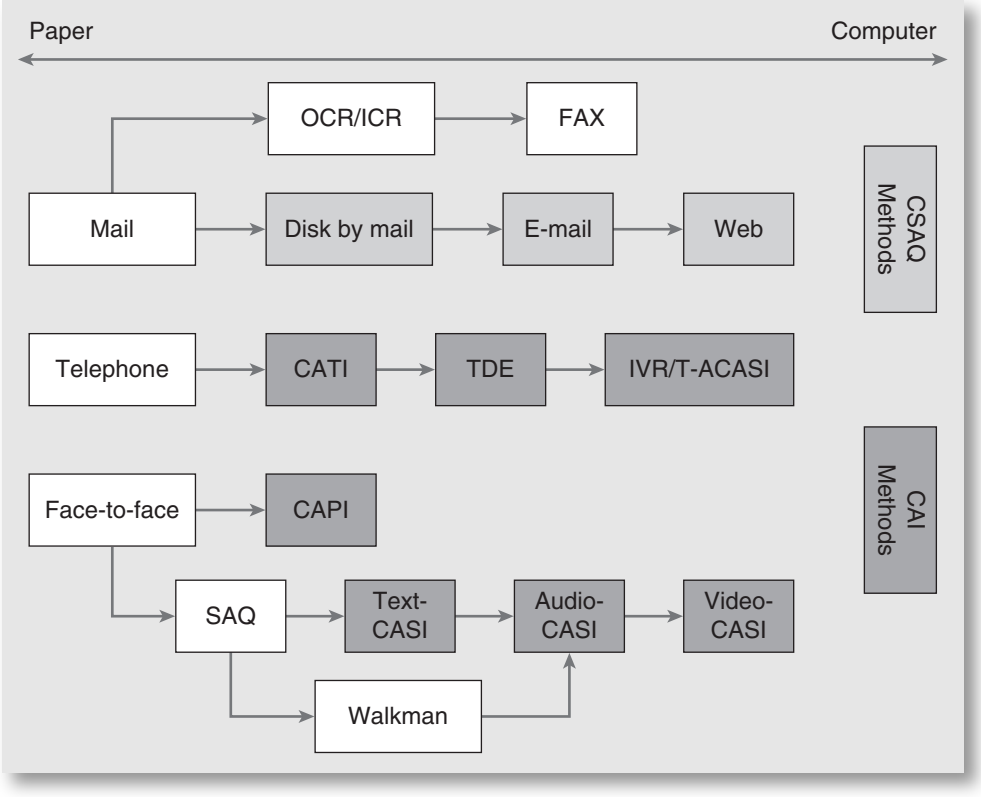
4. *What mode will be used to pose questions and collect answers from respondents?* That is to say, will respondents be questioned via a face-to-face interview, a phone call, a self-administered paper-and-pencil questionnaire, a diary, a form on the Internet, or some combination of these modes? Groves and colleagues (2004, p. 140) depict all these data collection modes in a historical perspective in a graph (Figure 2.2), with an accompanying glossary for its various acronyms.

Glossary of Acronyms

ACASI Audio computer-assisted self-interviewing. Here, the respondent operates the computer and enters his or her answers.

CSAQ Computerized self-administered questionnaires. These include questions posed both via e-mail and on a Web site.

Figure 2.2 The Evolution of Surveys



SOURCE: Groves et al. (2004, Figure 5.2, p. 140). Used by permission. Copyright John Wiley & Sons.

CAPI Computer-assisted personal interviewing. A computer displays questions on a screen, which an interviewer reads and then uses to enter the respondent's answer.

CASI Computer-assisted self-interviewing. Questions and responses are written (text-CASI), questions are presented in audio form (audio-CASI), or graphs are used to present questions (video-CASI).

CATI Computer-assisted telephone interviewing. This is the same as CAPI except that the interview is conducted over the phone.

OCR Optical character recognition, a technology that permits answers to mail questionnaires to be machine read. This technology was further advanced with the development of ICR, or "intelligent character recognition," which permits machines to read and code handwriting.

SAQ Self-administered questionnaires. Such questionnaires could be handed to or distributed to respondents in a group setting or mailed to them. This mode, for a brief time, used Walkman audio cassette tape players to ask questions, which were then answered on a paper form.

TDE Touchtone data entry. This mode is used to collect limited amounts of data by asking respondents to call a toll-free number and respond to recorded voice questions via the phone's keypad.

IVR Interactive voice response. A computer plays recordings of questions to a respondent over the phone, and responses are given either by using the phone's keypad or by answering aloud. This is also sometimes called T-ACASI because it's like ACASI but uses a phone.

Note that such a wide variety of modes of survey data collection and their combination make it difficult to draw any conclusions about which is "best" for a particular situation (although we will make some simple comparisons below). Indeed, no single mode is best for all circumstances. Choices must be made in the context of the objectives of the study and the personal, technological, and financial resources available.

5. *If interviewers are involved, how much effort will be devoted to training and supervising them?*

6. *How much effort will be devoted to checking the data files for accuracy and internal consistency?* An interviewer or respondent can miss a question or record a numeric response that falls outside the range of legitimate answers, or someone entering the data into a file (other than a respondent or an interviewer) may strike the wrong key. Questions that should have been skipped might have been answered and vice versa. In other words, the data may be "dirty." To what lengths will the collecting organization try to "clean" the data during the editing stage?

7. *What approaches will be used to adjust the survey estimates to correct for errors that can be identified?* Will the data be weighted to compensate for the discovery that certain groups of one's population are underrepresented among the final respondents (e.g., the poor may be less likely to have access to phones; those who work evenings and weekends may not be as likely to be at home when an interviewer calls). Will any adjustments be made for people who refuse or fail to answer an entire questionnaire (e.g., unit nonresponse) or individual questions in it (e.g., item nonresponse)? Will missing responses to individual questions cause such respondents to be excluded from any analysis that uses the answers to that

question, or will the values for these missing observations be imputed from knowledge of other information in the survey? Will measures discovered to lack sufficient **reliability** be adjusted and, if so, how?

Survey Mode. The range of different modes of sample surveys is quite large. Although it is difficult to make general assessments about these modes, research does suggest that some consequences follow from your choice of design and administration. One of the earliest such studies by Hochstim (1967) compared face-to-face, telephone, and mail modes of survey administration. The study randomly assigned households in Alameda County, California, to one of these three modes and to one of two different questionnaires. Like other studies, the face-to-face interviews produced the highest response rates but also had the highest costs. Telephone and mail questionnaires were within 12% of each other in cost. There were few substantive differences across the three modes, although mail questionnaires produced higher levels of some behaviors, such as reported alcohol consumption, and higher levels of nonresponse to specific questions, findings that have since been replicated.

A more recent study of the comparative costs of an RDD telephone survey versus an address-based mail questionnaire using the 2005 Behavioral Risk Factor Surveillance System showed that they were similarly close. The telephone version of the survey cost $79,578 per 1,000 completed interviews, 12% greater than the $70,969 per 1,000 completed mail questionnaires (which included a follow-up questionnaire mailing and a postcard reminder; Link, Battaglia, Frankel, Osborn, & Mokdad, 2008, p. 21).

Although there is no list of all individuals or all households in the United States from which one could draw a sample, the U.S. Postal Service (USPS) has created the computerized Delivery Sequence File (DSF), which contains all delivery-point addresses serviced by the USPS (except for general-delivery addresses). A number of studies (Iannacchione, Staab, & Redden, 2003; O'Muircheartaigh, Eckman, & Weiss, 2003) have demonstrated that the DSF may include up to 97% of all U.S. households. In the same methodological study noted above (Link et al., 2008), the address-based mail survey produced response rates higher than those of an RDD survey in five of the six states in which the study was conducted: 33.9% versus 29.4% in California, 39.9% versus 35.8% in Illinois, 26.2% versus 22.5% in New Jersey, 36.5% versus 31.1% in Texas, and 40.3% versus 34.1% in Washington. Phone interviews achieved a 45.8% response rate versus 37.0% for the mail version of the survey in North Carolina.

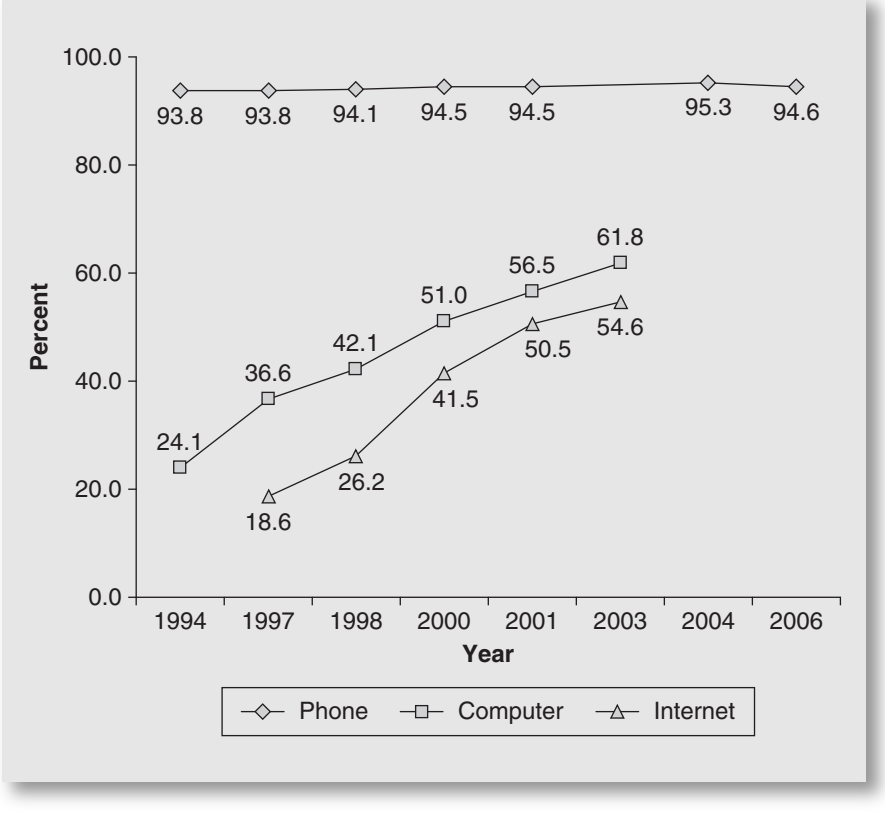
Of course, different modes have different levels of coverage; that is to say, they vary in how likely they are to reach a target audience. The combination of area probability sampling frames (a topic to be covered in Chapter 8) with

face-to-face interviews is typically the most likely to achieve the greatest level of coverage and completed interviews, although some trade-offs have to be made to minimize costs (e.g., members of the military, prisoners, or residents of Alaska and Hawaii are typically excluded from such samples because of the costs of accessing them).

Telephone coverage rates follow close behind, with computers and Internet access lagging behind in terms of the proportion of the U.S. public with access to these technologies, as displayed in Figure 2.3.

Note, however, that these numbers disguise differences in the prevalence of these technologies in U.S. households by important matters such as age, race, education, and income. While 96.6% of all U.S. households

Figure 2.3 Access to Phones, Computers, and the Internet in the United States



SOURCE: Bureau of Labor Statistics, Current Population Surveys. www.bls.gov/CPS

had access to a phone in 2006, only 87% of households with yearly income less than $10,000 did. In other words, a survey interested in the poor may potentially miss one of every eight such households if administered via telephone interviews, which is sometimes referred to as "coverage error." Similarly, while 54.6% of U.S households owned a computer in 2003, only 20% of households whose "head" did not complete high school owned one. On the other hand, 76.8% of households whose head had a bachelor's degree owned a computer.

Cell phones pose another challenge to research that relies on phone interviews because cell phones are often not included in telephone samples; a call to a cell can cost a respondent money. (This can be compensated by offering cell phone users proper remuneration for completing the survey.) Researchers will also not know whether respondents to cell phone calls are, say, driving a motor vehicle at the time the call is picked up, a condition that is illegal in some states and that has been shown to be associated with an increase in traffic accidents. Nor will a researcher know the environment in which the call is being taken. Those received while respondents are in public places may be less likely to get full and accurate responses. And the quality of transmissions through cell phones may also affect data quality: "Whatdyasay?"

Excluding cell phone numbers from phone surveys is increasingly difficult to do, however, because of the growing comingling of cell and landline numbers. Also, cell phone usage has become increasingly prevalent and thus hard to ignore. According to the 2006 National Health Interview Survey, 13% of U.S. households had only cell phones (Blumberg, 2007), a figure that Ehlen and Ehlen (2007) predict will soar to more than 40% of adults under the age of 30 (in contrast to less than 5% of adults 65 years of age and older) by 2009. These trends may well result in coverage errors if young adults are excluded from phone surveys in the future that rely entirely on sampling frames of landline phones. RDD phone surveys may now exclude up to 19% of all households in the United States (Link et al., 2008, p. 7). (For a general discussion of these and other issues that cell phones present in surveys, see Lavrakas, Shuttles, Steeh, & Fienberg, 2007.)

Finally, how do different data collection modes vary by costs? Not surprisingly, face-to-face interviews are typically the most costly, exceeding the costs of conducting interviews via phone by a factor of 2, although this ratio for national surveys may climb as high as 10 to 1 (Groves et al., 2004, p. 161). Phone interviews exceed the costs of mail questionnaires by only 1.2 to 1, although this varies by the number of callbacks, sample size, and so on.

No matter which mode of administration is used, sample surveys are often referred to as "observational" studies, in contrast to experimental designs. One of the most important weaknesses of such observational studies

is their inability to firmly establish causal relationships, which so much of our research would like to establish. This is so for several reasons. First, all the statistics discussed in this book either assume no causal relationship (but merely establish whether "high" values of one variable tend to be found to coexist with high values of one or more other variables) or assume that one or more variables cause another. The operative word here is *assume*, because it may well be the case that Y causes X instead of X causing Y. Research that collects data about possible treatments and outcomes simultaneously is hard-pressed to demonstrate causation conclusively.

The joint association between two variables may, for example, be the **spurious** consequence of their shared association with a third variable, and they may therefore not be causally related at all. Perhaps the most famous instance of such a relationship is that found between ice cream consumption and crime. Although there is an empirical relationship between the two, one would be hard-pressed to argue that ice cream consumption causes crime. Indeed, their association appears to be the result of their joint occurrence with warm weather, which appears to cause, encourage, or enable both crime rates and ice cream consumption to rise.

It is also the case that observational studies don't introduce a treatment and assess its consequences but, rather, observe a consequence and then attempt to determine whether that consequence (e.g., unemployment) tends to vary with other measured characteristics (e.g., criminal record, education levels). To discover, however, that men who have been incarcerated have poor employment outcomes could well be the spurious consequence of characteristics of these men (e.g., alcohol/drug abuse, behavioral problems, poor interpersonal skills) that are likely to "cause" both poor employment outcomes and higher rates of incarceration (see, e.g., Pager, 2003). Similarly, it is difficult to conclude that participation in job training programs "causes" better employment outcomes by observing higher proportions of such participants with jobs than people who didn't participate in these training programs. Those with better employment opportunities may have self-selected into such job training programs and would have found employment even without having completed a training program (see, e.g., Winship & Morgan, 1999).

The final distinction to be made about surveys is that between cross-sectional and longitudinal surveys. Cross-sectional surveys measure the attitudes and behaviors of individuals at one point in time. Longitudinal or panel studies collect information about individuals (or whatever type of unit you're studying) repeatedly—that is to say, two or more times. A modest confusion may arise here because you can string together a series of repeated cross-sectional surveys that may look like a longitudinal study because it may be, say, responses of adults in the United States in 1970, 1971, 1972, and so on.

The distinction between cross-sectional and longitudinal designs is important for at least two reasons. First, as we will see again later in this book, observations in longitudinal studies are said to be "dependent." A response to a questionnaire in Year 2 in such a study is a function of being a respondent in Year 1 of that study. Dependent observations—in this particular meaning of the term *dependent* (we'll see other meanings later)—require different statistical tools or tests from those observations (such as those in cross-sectional surveys) that are said to be "independent" of one another. My random selection to participate in a cross-sectional survey has nothing to do with your random selection to participate. Thus, the observations in a cross-sectional survey are said to be independent. Second, longitudinal surveys—largely because of repeated measures over time—can provide somewhat more leverage in drawing cause-and-effect conclusions than can cross-sectional surveys (see, e.g., Singer & Willett, 2003). (Structural equation models, which are beyond the scope of this book, were developed to remedy this shortcoming of cross-sectional surveys, but they require some assumptions that can be difficult to defend.)

Administrative Records

Organizations collect data about themselves. Sometimes, a lot of data. They are often required to do so. Financial data, for example, are required by an organization's auditors and the Internal Revenue Service. Budgets are required for nearly all of an organization's planning processes. Universities are required to report crimes committed on campus. And many organizations collect data to assess and evaluate whether they are achieving their mission, although this is not always easy or inexpensive to do.

The types of administrative records are as diverse as the organizations themselves. They include, but are by no means limited to, orders, invoices, payments, receipts, insurance costs, the number of clients served, the number and level of donations received, employee data concerning employee backgrounds and absenteeism rates, fines and fees, customer complaints, and complaints responded to within specified periods of time.

This type of data has many of the same problems with quality and completeness as sample surveys. They are also not likely to include all the information one would like to have in order to test for a full range of possible explanations—say, for differences in the level of your employees' performance.

Administrative records, however, can be nonetheless indispensable for answering many questions about organizational performance. They can be made even more useful when combined with data explicitly drawn from a

sample of an organization's members. Such information may also be strung together as part of a time series, say, of traffic fatalities in a state in a specified number of years. Such data can be used—for example, to assess the effects of changes in law enforcement practices (e.g., a crackdown on speeding violations, revocation of a driver's license for speeding)—in what is called an interrupted time-series analysis, to which we'll return in Chapter 13.

"Actors," "Confederates," "Testers," and "Audits"

Some research hires people to play a role that is intended to evoke a response that can be compared with responses to, say, different types of "actors." This type of research was first launched in a major, systematic way in studies by the U.S. Department of Housing and Urban Development to test housing discrimination (Wienk et al., 1979). The design has also been used, for example, to study the presence and consequences of racial stereotyping and discrimination in job markets.

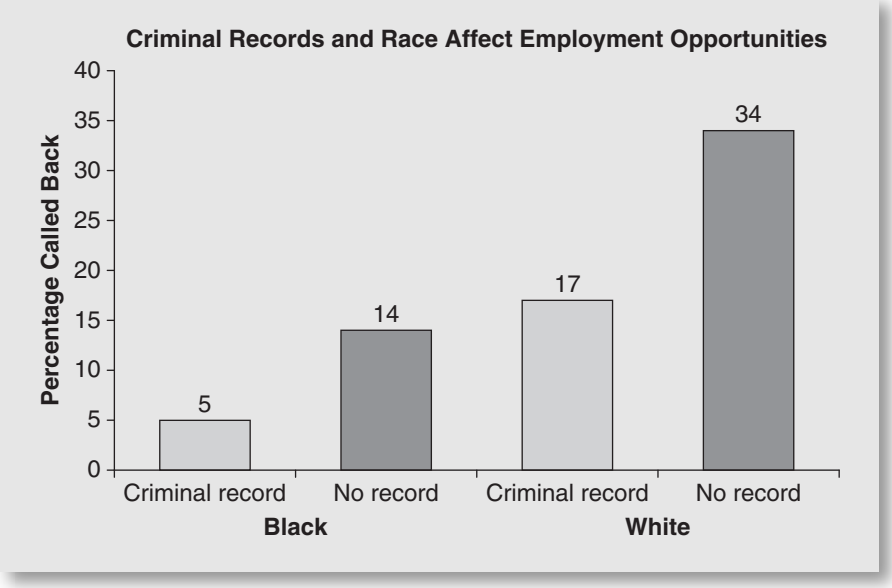
One example of this type of study sought to understand the effects of incarceration on job opportunities and outcomes by conducting an experiment in Milwaukee between June and December 2001 (Pager, 2003). Pairs of white and black actors, also known as "testers" or "auditors," submitted applications for entry-level jobs advertised in a local newspaper and a state Web site for employment listings. Each member of the pair was randomly assigned to play the role of someone with a past felony conviction and 18-month prison sentence applying to randomly selected job listings. The testers were matched as best they could on personal qualities (although the vast majority of applications [76%] included no face-to-face interview, which made the possible differences between testers less important). Their applications differed otherwise only by race and criminal record.

The results were fairly dramatic if not unexpected, as represented in Figure 2.4 (from Pager, 2003, Figure 6, p. 958). The bars represent the percentage of these different categories of applicant who were called (in some cases, called back) for an interview. Both race and previous incarceration made a big difference in whether someone was asked for an interview.

Audits that combine experimental designs with "real-world" field settings are a powerful tool for uncovering actual behavior patterns, in contrast to intentions, expectations, attitudes, or opinions, on which many sample surveys focus.

There are, however, several potential shortcomings in such a design. First, the testers may not be precisely matched on the characteristics the investigator would like to hold constant, although alternating testers to take on different

Figure 2.4 Criminal Records, Race, and Employment Opportunities



SOURCE: Pager (2003, Figure 6, p. 958).

roles may well mitigate this problem. You cannot, however, easily have an actor change race from one setting to another. Second, audits are usually carried out in a specific location (e.g., Milwaukee) or small set of locations, thus raising the question of whether those locales are similar to or differ from the larger set of locations about which one would typically like to make inferences. Third, the testers, no matter how well rehearsed, may influence the results in nonrandom ways. It may be difficult to play the role of a drug felon in a job interview, and that discomfort might be communicated in such a way as to affect the likelihood of a callback. Finally, such designs are predicated on a lie or deceit in some way or another, ranging from the actors lying about their own backgrounds to the failure to disclose the fact that the activities are part of a research project that could potentially put the subjects at some risk of violating the law (see Heckman & Seligman, 1993, for a critique of this method).

The merits of this form of research design are apparent, however. These designs have been used to study questions that avoid the "social desirability" effect, where respondents give responses to place themselves favorably in the eyes of the questioner. This problem often troubles sample surveys as well as other study designs in which a respondent/participant is aware of being observed.

Other Observational Studies (Unobtrusive and Participant)

Two more types of studies fall under the rubric of "observational" studies. The first is illustrated by a study that seeks to learn whether sports utility vehicle (SUV) drivers drive more aggressively than drivers of smaller automobiles. Observers would be instructed on what types of behaviors constituted "aggressive behavior" in the eyes of the principal investigators. The observers would position themselves on street corners and overpasses and look for the designated types of behavior in some sample of cars and SUVs that passed their observation points at designated (and randomly selected) times. Obviously, SUV and other automobile drivers are not being asked whether they drive aggressively. Their actual behavior is being observed instead (unobtrusively, you might say).

Although this type of study cannot be used to study all questions, it has the advantage of observing and measuring actual behavior instead of attitudes or intentions, which are only imperfectly related to behavior. Some attitudes and intentions are weakly held. Barriers and incentives can overpower them. "The road to hell is paved with good intentions," as the old proverb goes.

The second type of observational study embeds observers among the populations that are the object of study. This technique is commonplace in anthropology. It is often called participant observation or ethnography. Its methods were made famous by studies such as Margaret Mead's work with the Trobriand Islanders, but the method retains strong advocates among a wide range of researchers who study a variety of questions (see, e.g., Anderson et al., 2004). The method's strengths lie in the ability of observers to understand how others give meaning to their circumstances and interpret their realities and why others think and behave as they do.

The depth of meaning that these methods are capable of exploring comes at the expense of not being able to make estimates with a certain degree of certainty and precision about the prevalence of conditions in a population or the strength of the relationship among two or more concepts or differences between two or more groups. And some participant observers become so much a part of the population they're studying (i.e., "go native") as to color their ability to make unbiased observations.

Focus Groups

Focus groups are commonplace in market research. They typically involve the gathering of about a dozen people in relatively homogeneous

groups who engage in semistructured, but highly open-ended, conversations. They are often used to test reactions to a new product or idea and, like ethnographic methods, to better understand how and why people hold certain attitudes or beliefs. One of the strengths of this method lies in the opportunity it provides to watch and observe nonverbal reactions as well as verbal responses. It can explore new ideas without implementing them. The group settings can trigger more ideas than a one-on-one interview. And this method can be used—like ethnographic research—to learn what questions to ask through more quantitative techniques (e.g., a sample survey).

Focus groups also have several potential and real weaknesses. Strong personalities can sway others' opinions (although this can be guarded against by a good facilitator). People may want to tell you what you want to hear (although this **social desirability effect** is present in many of the different research designs described in this chapter). And they cannot provide estimates about the prevalence of conditions in a population or the strength of the relationship among two or more concepts or differences between two or more groups with any degree of certainty and precision.

● NONRESPONSE MAY (OR MAY NOT) LEAD TO NONRESPONSE BIAS

Surveys of humans rarely, if ever, obtain completed interviews (questionnaires) from every one who is sampled. Mail questionnaires to randomly selected members of lists often have response rates of no more than 5%. Telephone surveys such as the University of Michigan's Survey of Consumers, which measure consumer attitudes and expectations, complete interviews with only about 60% of the selected adults, and about 17% of sampled schools in the National Assessment of Educational Progress refuse to participate in the tests that produce "The Nation's Report Card." In general, face-to-face surveys produce the highest response rates, phone surveys the second, and mail questionnaires the lowest (Hox & de Leeuw, 1994), although Link and colleagues (2008) demonstrate a slight edge to an address-based mail questionnaire over an RDD survey. Among self-administered questionnaires, paper-based formats tend to outperform e-mail and Web-based formats (Couper, 2001). Nonresponse and refusal rates have risen steadily across many ongoing surveys over time.

Survey researchers have developed and tested a number of tactics for reducing nonresponse, including the following:

- Repeated callbacks in face-to-face and phone interviews and repeated follow-ups in self-administered questionnaires (The Adult Education and Lifelong Learning [AELL] survey of the U.S. Department of Education, for example, used up to 20 call attempts to complete each interview in its 2001 survey.)
- Letters in advance of the interviewer contact
- Refusal conversion letters or phone calls, often from especially persuasive interviewers, to convert refusals into completed interviews

Methodological experiments using these techniques in AELL 2001 and a 2003 Pew Research Center Survey (both of which were telephone surveys) increased response rates from 34% to 43.4% and from 27.7% to 51.4%, respectively (Montaquila, Brick, Hagedorn, Kennedy, & Keeter, 2008, p. 576).

Whether low response rates cause a problem for the data analyst is predicated on the differences between respondents and nonrespondents. Do these two groups differ in their socio-demographic makeup, attitudes, or behavior? If they do, the analyst should be wary about drawing conclusions about the population from which the sample was drawn.

Interestingly, the AELL 2001 and 2003 Pew Research Center surveys noted above demonstrated few substantive differences between data sets based on minimal and extensive efforts to increase response rates.

It is often the case that we do not know about the attitudes and behaviors of a study's nonrespondents. It is more often the case, however, that we have independent data about the socio-demographic characteristics of the populations being sampled. The age, race, and gender distribution, say, of Orange County, Florida, residents are available from the U.S. Census Bureau and can be compared, for example, with the age, race, and gender distribution of any sample drawn from that population.

MULTIMETHOD DESIGNS AND CONVERGENT RESULTS CAN OVERCOME THE LIMITATIONS OF ANY SINGLE RESEARCH DESIGN ●

Different research designs have different strengths and weaknesses. Some are better for answering some types of questions. The choice of the most appropriate research design depends on your questions, as well as the resources and capacities you (or others you hire) have to carry out the study. It should also be clear that research that combines several different designs is more likely to bear fruitful and insightful results and conclusions than research that

relies on a single data collection strategy. The more viewpoints you bring to bear, the more likely you are to see the whole and rich complexity of the answers you seek. Even advocates of randomized control studies caution strongly against making generalizations to people or organizations beyond any particular experiment (Shadish & Cook, 1999, p. 299).[3]

The important point to repeat here is that the more the studies that address your questions, the more confident you can be in their answers and the decisions you base on them (assuming, of course, that the different studies arrive at the same conclusions). Such convergent results, for example, led C. Everett Koop as Surgeon General to conclude that cigarette smoking caused lung cancer despite the absence of RCTs (experiments) on that question. That is to say, people were not randomly assigned to smoke the equivalent of a pack a day for 20 years and compared with another group of people randomly prohibited from smoking or inhaling smoke from others. Yet the evidence from multiple studies and the size of the effect across many of these studies were collectively persuasive enough for most to draw the causal connection between smoking and lung cancer. Executives of some tobacco companies argued for years that the correlation found between cigarette smoking and cancer was the result of a hereditary condition that caused both cancer and a desire to smoke cigarettes. That is to say, they argued that the relationship between smoking and lung cancer was "spurious" and that extant studies (none of which were RCTs) had failed to rule out alternative causal explanations or the hereditary factors that were not measured in these studies. We will see later that this type of critique—"You didn't include a measure for a plausible alternative explanation in your study!"—is one that can be levied against nearly any nonexperimental study. Such a charge falls under the rubric of "failure to fully specify your model." It's a critique to anticipate and guard against in your own research (if possible).

● CONCLUSION

We turn in the next chapter to the concern of how best to ask questions, whatever the research design. Some of you may find it odd that I'm paying so much attention to matters of research design and question and questionnaire construction in a textbook on statistics. To repeat a metaphor, statistics are tools with which to analyze data derived from one or more data collection methods, each with its own strengths and weaknesses. These designs can, however, produce seriously flawed data, inadequate to provide answers to the questions at hand. If so, no statistical technique—no matter how sophisticated—can save the study. No decisions can be confidently made

with poor data, no building confidently constructed from faulty building materials and poor architectural plans. Or, as Light, Singer, and Willett write (1990), "You can't fix by analysis what you bungled by design" (p. v).

Richard Berk (2004, pp. 234–237) illustrates this point forcefully in describing the flaws in a study of racial profiling for the Los Angeles City Police Department (LAPD) in 2000–2001. The story begins with the mayor of Los Angeles signing a consent decree between the city and the U.S. Department of Justice in response to charges of corruption and misconduct. The decree required the Los Angeles police to provide quarterly "audits" of specific police activities to the local police commission and inspector general.

The audit reports were well staffed and funded. The LAPD hired a technical consultant to collect and organize the required data. The reports were also closely scrutinized by the mayor, the police commission, the city council, the police union, and the media.

The LAPD had been charged specifically with racial profiling of pedestrians and motorists. The consent decree sought to assess the veracity of this charge, which required the collection of information about the race and ethnicity of all people whom the police stopped and/or detained. Police officers were required to record on new forms the following information for each person stopped (Berk, 2004, p. 42):

1. Whether the individual was a pedestrian, a driver, or a passenger
2. Gender
3. Apparent descent (racial/ethnic background)
4. Apparent age
5. Incident number
6. The initial reason for being stopped
7. Whether the driver was required to exit the vehicle
8. Whether a patdown or frisk was required
9. Whether the detainee was asked to submit to a consensual search
10. If there was a warrantless search, the search authority
11. Whether a search was actually conducted
12. What was searched
13. What was discovered/seized
14. The action taken

15. Date
16. Time
17. Reporting district
18. Officer's serial number (for two officers if necessary)
19. Officer's police division number (for two officers if necessary)
20. Officer's name (for two officers if necessary)

Unfortunately, no information was collected about the location of the stop. No effort was made to collect baseline information with which to compare the data that were recorded on these new forms. Assuming that police officers could accurately observe and record race, ethnicity, and age, they were not asked to report the relative mix of motorists' or pedestrians' characteristics at the location where each stop occurred. Such data could have been collected through any number of means. Moreover, no outcome data were recorded. For example, what proportion of stops led to an arrest and conviction? Stops of minorities that led to no such outcomes could surely be suspected of racial profiling. And how do these outcome data compare with data from police departments in other cities?

Professor Berk (2004) concludes this story as follows:

> Clearly, important data are not being collected. And the data that ultimately will be available will have significant problems. One can predict, nevertheless, that there will be hundreds of pages of regression output addressing racial profiling. Reports from those analyses will be laden with *p*-values, hypothesis tests, and lots of causal talk. (p. 237)

You too will learn about *p* values, hypothesis tests, and more. You will even produce pages (not hundreds) of regression output if you complete the exercises that accompany most of the chapters of this book. I hope that you will do so with an understanding of the appropriate use of specific statistics under different conditions. But we must first understand what data to collect, what comparisons will help answer our questions, and what questions to ask and how.

● NOTES

1. Some of these experiments include the added feature of making assignments to treatment and control groups unknown and unknowable by the investigators as

well as study subjects, to eliminate any contamination of outcomes by the experimenter himself. These RCTs are called "double-blind" as a consequence.

2. For a state-by-state listing of academic and not-for-profit survey research organizations, see Bradburn, Sudman, and Wansink (2004, Appendix A).

3. Indeed, meta-analysis was developed about 20 years ago to systematically bring together multiple studies on the same question. This is a topic worthy of your attention but beyond the reach of this book. For those interested in exploring meta-analysis, consult Lipsey and Wilson (2001).

No exercise here. One awaits the conclusion of the next chapter. Keep reading.