## CHAPTER NINE

# Survey Research

PENNY S. VISSER, JON A. KROSNICK, AND PAUL J. LAVRAKAS

Social psychologists have long recognized that every method of scientific inquiry is subject to limitations and that choosing among research methods inherently involves trade-offs. With the control of a laboratory experiment, for example, comes an artificiality that raises questions about the generalizability of results. And yet the "naturalness" of a field study or an observational study can jeopardize the validity of causal inferences. The inevitability of such limitations has led many methodologists to advocate the use of multiple methods and to insist that substantive conclusions can be most confidently derived by triangulating across measures and methods that have nonoverlapping strengths and weaknesses (see, e.g., Brewer, this volume, Ch. 1; Campbell & Fiske, 1959; Campbell & Stanley, 1963; Cook & Campbell, 1969; Crano & Brewer, 1986; E. Smith, this volume, Ch. 2).

This chapter describes a research methodology that we believe has much to offer social psychologists interested in a multimethod approach: survey research. *Survey research* is a specific type of field study that involves the collection of data from a sample of elements (e.g., adult women) drawn from a well-defined population (e.g., all adult women living in the United States) through the use of a questionnaire (for more lengthy discussions, see Babbie, 1990; Fowler, 1988; Frey, 1989; Lavrakas, 1993; Weisberg, Krosnick, & Bowen, 1996). We begin the chapter by suggesting why survey research may be valuable to social psychologists and then outline the utility of various study designs. Next, we review the basics of survey sampling and questionnaire design. Finally, we describe procedures for pretesting questionnaires and for data collection.

## REASONS FOR SOCIAL PSYCHOLOGISTS TO CONDUCT SURVEY RESEARCH

Social psychologists are interested in understanding how people influence, and are influenced by, their social environment. And to the extent that social psychological phenomena are universal across different types of people, it makes little difference precisely with whom social psychological research is conducted – even data collected from samples that are decidedly unrepresentative of the general population can be used to draw inferences about that population.

In recent years, however, psychologists have become increasingly sensitive to the impact of dispositional and contextual factors on human thought and social behavior. Instead of broad statements about universal processes, social psychologists today are far more likely to offer qualified accounts of which people, under which conditions, are likely to exhibit a particular psychological phenomenon or process. And accordingly, social psychologists have increasingly turned their attention to interactions between various social psychological processes and characteristics of the individual, such as personality traits, identification with a social group or category, or membership in a distinct culture. In many cases, the nature of basic social psychological processes has been shown to depend to a large degree on characteristics of the individual.

This chapter was completed while the second author was a Fellow at the Center for Advanced Study in the Behavioral Sciences, supported by National Science Foundation Grant SBR-9022192. Correspondence should be directed to Penny S. Visser, Department of Psychology, Princeton University, Princeton, New Jersey 08544 (e-mail: pvisser@Princeton.edu), or Jon A. Krosnick, Department of Psychology, Ohio State University, 1885 Neil Avenue, Columbus, Ohio 43210 (e-mail: Krosnick@osu.edu).

The process by which attitude change occurs, for example, has been shown to differ for people who are low and high in what Petty and Cacioppo (1986) have termed "need for cognition," a general enjoyment of and preference for effortful thinking. Attitude change among people high in need for cognition tends to be mediated by careful scrutiny of the arguments in a persuasive appeal, whereas attitude change among people low in need for cognition tends to be based on cues in the persuasive message or context, such as the attractiveness of the source.

Similarly, attributions have been shown to differ depending on social group membership (see, e.g., Hewstone, Bond, & Wan, 1983). People tend to attribute positive behaviors by members of their own social group or category to stable, internal causes. Those same positive behaviors performed by a member of a different social group, however, are more likely to be attributed to transitory or external factors.

According to much recent research, culture may also moderate many social psychological phenomena. Markus and Kitayama (1991), for example, have argued that members of different cultures have different construals of the self and that these differences can have a profound impact on the nature of cognitive, emotional, and motivational processes. Similarly, Nisbett and his colleagues (Cohen, Nisbett, Bowdle, & Schwarz, 1996; Nisbett, 1993; Nisbett & Cohen, 1996) have explored what they termed the "culture of honor" of the American South and have demonstrated marked differences in the cognitive, emotional, behavioral, and even physiological reactions of southern men (relative to their northern counterparts) when confronted with insult.

These kinds of process-by-individual-difference interactions suggest that precisely who participates in social psychological research can have a profound impact on what results are obtained. And of course, for the vast majority of social psychological research, that "who" has been the infamous college sophomore. Sears (1986) has argued that the field's overwhelming reliance on this narrow base of research participants may represent a serious problem for social psychology. Pointing to various attributes that are characteristic of young adults, Sears (1986) suggested that the "college sophomore" participant pool is unrepresentative of the general population in a number of important ways. Among other things, young adults are more susceptible to attitude change (Alwin, Cohen, & Newcomb, 1991; Glenn, 1980; Krosnick & Alwin, 1989; Sears, 1983), exhibit less stable personality traits (Caspi, Bem, & Elder, 1989; Costa, McCrae, & Arenberg, 1983; Nesselroade & Baltes, 1974), have more weakly established self-images (Mortimer, Finch, & Kumka, 1982), and have less well-developed social identities (Alwin et al., 1991) than·do older adults.

Because of these kinds of differences, Sears (1986) argued, the field's reliance on participant pools of college-aged adults raises questions about the generalizability of some findings from social psychological laboratory research and may have contributed to a distorted portrait of "human nature." However, the evidence Sears (1986) cited largely reveals the prevalence of certain characteristics (e.g., the frequency of attitude change or the firmness of social identities), rather than differences in the processes by which these characteristics or others emerge in different age groups. We currently know so little about the operation of social psychological processes in other subsets of the population that it is impossible to assess the extent of bias in this regard.

Doing so will require studies of samples that are representative of the general population, and inducing most members of such samples to visit a laboratory seems practically impossible. Studying a representative sample through field research, however, is relatively easy and surprisingly practical. Using the basic tenets of probability theory, survey researchers have developed a number of efficient strategies for drawing representative samples that are easy to contact. And when samples have been selected in such a manner, social psychologists can confidently generalize findings to the entire population. Furthermore, survey research provides ideal conditions for the exploration of Process × Individual Difference interactions because carefully selected samples reflect the full heterogeneity of the general population.

There are two primary limitations of survey research for social psychologists. First, surveys are more expensive and time-consuming than most laboratory experiments using captive participant pools. However, many cost-saving approaches can be implemented. Second is the impracticality of executing elaborate scripted scenarios for social interaction, especially ones involving deception. Whereas these sorts of events can be created in labs with undergraduate participants, they are tougher to do in the field. But as we discuss shortly, many experimental procedures and manipulations can be incorporated in surveys.

Put simply, social psychology can happily proceed doing most of our research with college sophomores, assuming that our findings generalize. And we can live with the skepticism of scholars from other disciplines who question that generalizability, having documented the profound impact that context and history have on social processes. Or we can accept the challenge and explore the replicability of our findings in the general population. Either we will confirm our assumptions of generalizability or we will refine our theories by adding to them new mediators and moderators. The explication of the survey method offered below is intended to help those who accept the challenge.

## TOTAL SURVEY ERROR

Even researchers who recognize the value of survey methodology for social psychological inquiry are sometimes reluctant to initiate survey research because of misconceptions regarding the feasibility of conducting a survey on a limited budget. And indeed, the cost of prominent large-scale national surveys conducted by major survey organizations are well outside of the research budgets of most social psychologists. But survey methodologists have recently begun to rekindle and expand the early work of Hansen and his colleagues (e.g., Hansen & Madow, 1953) in thinking about survey design issues within an explicit cost–benefit framework geared toward helping researchers make design decisions that maximize data quality within the constraints of a limited budget. This approach to survey methodology, known as the "total survey error" perspective (cf. Dillman, 1978, Fowler, 1988; Groves, 1989; Lavrakas, 1993), can provide social psychologists with a broad framework and specific guidelines for making decisions to conduct good surveys on limited budgets while maximizing data quality.

The total survey error perspective recognizes that the ultimate goal of survey research is to accurately measure particular constructs within a sample of people who represent the population of interest. In any given survey, the overall deviation from this ideal is the cumulative result of several sources of survey error. Specifically, the total survey error perspective disaggregates overall error into four components: coverage error, sampling error, nonresponse error, and measurement error. *Coverage error* refers to the bias that can result when the pool of potential survey participants from which a sample is selected does not include some portions of the population of interest. *Sampling error* refers to the random differences that invariably exist between any sample and the population from which it was selected. *Nonresponse error* is the bias that can result when data are not collected from all of the members of a sample. And *measurement error* refers to all distortions in the assessment of the construct of interest, including systematic biases and random variance that can be brought about by respondents' own behavior (e.g., misreporting true attitudes, failing to pay close attention to a question), interviewer behavior (e.g., misrecording responses, providing cues that lead participants to respond in one way or another), and the questionnaire (e.g., ambiguous or confusing question wording, biased question wording or response options).

The total survey error perspective advocates explicitly taking into consideration each of these sources of error and making decisions about the allocation of finite resources with the goal of reducing the sum of the four. In the sections that follow, we consider each of these potential sources of survey error and their implications for psychologists seeking to balance pragmatic budget considerations against concerns about data quality.

## STUDY DESIGNS

Surveys offer the opportunity to execute studies with various designs, each of which is suitable for addressing particular research questions of long-standing interest to social psychologists. In this section, we will review several standard designs, including cross-sectional, repeated cross-sectional, panel, and mixed designs, and discuss when each is appropriate for social psychological investigation. We will also review the incorporation of experiments within surveys.

### Cross-Sectional Surveys

Cross-sectional surveys involve the collection of data at a single point in time from a sample drawn from a specified population. This design is most often used to document the prevalence of particular characteristics in a population. For example, cross-sectional surveys are routinely conducted to assess the frequency with which people perform certain behaviors or the number of people who hold particular attitudes or beliefs. However, documenting prevalence is typically of little interest to social psychologists, who are usually more interested in documenting associations between variables and the causal processes that give rise to those associations.

Cross-sectional surveys do offer the opportunity to assess relations between variables and differences between subgroups in a population. But although many scholars believe their value ends there, this is not the case. Cross-sectional data can be used to test causal hypotheses in a number of ways. For example, using statistical techniques such as two-stage least squares regression, it is possible to estimate the causal impact of variable A on variable B, as well as the effect of variable B on variable A (Blalock, 1972). Such an analysis rests on important assumptions about causal relations among variables, but these assumptions can be tested and revised as necessary (see, e.g., James & Singh, 1978). Furthermore, path analytic techniques can be applied to test hypotheses about the mediators of causal relations (Baron & Kenny, 1986; Kenny, 1979), thereby validating or challenging notions of the psychological mechanisms involved. And cross-sectional data can be used to identify the moderators of relations between variables, thereby also shedding some light on the causal processes at work (e.g., Krosnick, 1988b).

A single, cross-sectional survey can even be used to assess the impact of a social event. For example, Krosnick and Kinder (1990) studied priming in a real-world setting by focusing on the Iran/Contra scandal. On November 25, 1986, the American public learned that members of the National Security Council had been funneling funds (earned through arms sales to Iran) to the Contras fighting to overthrow the Sandinista government in Nicaragua. Although there had been almost no national news media attention to Nicaragua and the Contras previously, this revelation led to a dramatic increase in the salience of that country in the American press during the following weeks. Krosnick and Kinder suspected that this coverage might have primed Americans' attitudes toward U.S. involvement in Nicaragua and thereby increased the impact of these attitudes on evaluations of President Ronald Reagan's job performance.

To test this hypothesis, Krosnick and Kinder (1990) took advantage of the fact that data collection for the 1986 National Election Study, a national survey, was underway well before November 25 and continued well after that date. So these investigators simply split the survey sample into one group of respondents who had been interviewed before November 25 and the others, who had been interviewed afterward. As expected, overall assessments of presidential job performance were based much more strongly on attitudes toward U.S. involvement in Nicaragua in the second group than in the first group.

Furthermore, Krosnick and Kinder (1990) found that this priming effect was concentrated primarily among people who were not especially knowledgeable about politics (so-called "political novices"), a finding permitted by the heterogeneity in political expertise in a national sample of adults. From a psychological viewpoint, this suggests that news media priming occurs most when opinions and opinion-formation processes are not firmly grounded in past experience and in supporting knowledge bases. From a political viewpoint, this finding suggests that news media priming may not be especially politically consequential in nations where political expertise is high throughout the population.

### Repeated Cross-Sectional Surveys

Additional evidence consistent with a hypothesized causal relation would be that changes over time in a dependent variable parallel changes in a proposed independent variable. One way to generate such evidence is to conduct repeated cross-sectional surveys, in which data are collected from independent samples drawn from the same population at two or more points in time. If a hypothesized causal relation exists between two variables, between-wave changes in the independent variable should be mirrored by between-wave changes in the dependent variable. For example, if one believes that interracial contact may reduce interracial prejudice, an increase in interracial contact over a period of years in a society should be paralleled by a reduction in interracial prejudice.

One study along these lines was reported by Schuman, Steeh, and Bobo (1985). Using cross-sectional surveys conducted between the 1940s and the 1980s in the United States, these investigators documented dramatic increases in the prevalence of positive attitudes toward principles of equal treatment of Whites and Blacks. And there was every reason to believe that these general principles might be important determinants of people's attitudes toward specific government efforts to ensure equality. However, there was almost no shift during these years in public attitudes toward specific implementation strategies. This challenges the notion that the latter attitudes were shaped powerfully by the general principles.

Repeated cross-sectional surveys can also be used to study the impact of social events that occurred between the surveys (e.g., Weisberg, Haynes, & Krosnick, 1995). And repeated cross-sectional surveys can be combined into a single data set for statistical analysis, using information from one survey to estimate parameters in another survey (e.g., Brehm & Rahn, 1997).

### Panel Surveys

In a panel survey, data are collected from the same people at two or more points in time. Perhaps the most obvious use of panel data is to assess the stability of psychological constructs and to identify the determinants of stability (e.g., Krosnick, 1988a; Krosnick & Alwin, 1989). But with such data, one can test causal hypotheses in at least two ways. First, one can examine whether individual-level changes over time in an independent variable correspond to individual-level changes in a dependent variable over the same period of time. So, for example, one can ask whether people who experienced increasing interracial contact manifested decreasing racial prejudice, while at the same time, people who experienced decreasing interracial contact manifested increasing racial prejudice.

Second, one can assess whether changes over time in a dependent variable can be predicted by prior levels of an independent variable. So, for example, do people who had the highest amounts of interracial contact at Time 1 manifest the largest decreases in racial prejudice between Time 1 and Time 2. Such a demonstration provides relatively strong evidence consistent with a causal hypothesis, because the changes in the dependent variable could not have caused the prior levels of the independent variable (see, e.g., Blalock, 1985; Kessler & Greenberg, 1981, on the methods; see Rahn, Krosnick, & Breuning, 1994, for an illustration of its application).

One application of this approach occurred in a study of a long-standing social psychological idea called the *projection hypothesis*. Rooted in cognitive consistency theories, it proposes that people may overestimate the extent to which they agree with others whom they like, and they may overestimate the extent to which they disagree with others whom they dislike. By the late 1980s, a number of cross-sectional studies by political psychologists yielded correlations consistent with the notion that people's perceptions of the policy stands of presidential candidates were distorted to be consistent with attitudes toward the candidates (e.g., Granberg, 1985; Kinder, 1978). However, there were alternative theoretical interpretations of these correlations, so an analysis using panel survey data seemed in order. Krosnick (1991a) did just such an analysis exploring whether attitudes toward candidates measured at one time point could predict subsequent shifts in perceptions of presidential candidates' issue stands. And he found no projection at all to have occurred, thereby suggesting that the previously documented correlations were more likely due to other processes (e.g., deciding how much to like a candidate based on agreement with him or her on policy issues; see Byrne, 1971; Krosnick, 1988b).

The impact of social events can be gauged especially powerfully with panel data. For example, Krosnick and Brannon (1993) studied news media priming using such data. Their interest was in the impact of the Gulf War on the ingredients of public evaluations of presidential job performance. For the 1990–1991 National Election Panel Study of the Political Consequences of War, a panel of respondents had been interviewed first in late 1990 (before the Gulf War) and then again in mid-1991 (after the war). The war brought with it tremendous news coverage of events in the Gulf, and Krosnick and Brannon suspected that this coverage might have primed attitudes toward the Gulf War, thereby increasing their impact on public evaluations of President George Bush's job performance. This hypothesis was confirmed by comparing the determinants of presidential evaluations in 1990 and 1991. Because the same people had been interviewed on both occasions, this demonstration is not vulnerable to a possible alternative explanation of the Krosnick and Kinder (1990) results described above: that different sorts of people were interviewed before and after the Iran/Contra revelation and their preestablished presidential evaluation strategies may have produced the patterns of regression coefficients that would then have been misdiagnosed as evidence of news media priming.

Panel surveys do have some disadvantages. First, although people are often quite willing to participate in a single cross-sectional survey, fewer may be willing to complete multiple interviews. Furthermore, with each additional wave of panel data collection, it becomes increasingly difficult to locate respondents to reinterview them, because some people move to different locations, some die, and so on. This may threaten the representativeness of panel survey samples if the members of the first-wave sample who agree to participate in several waves of data collection differ in meaningful ways from the people who are interviewed initially but do not agree to participate in subsequent waves of interviewing.

Also, participation in the initial survey may sensitize respondents to the issues under investigation, thus changing the phenomena being studied. As a result, respondents may give special attention or thought to these issues, which may have an impact on subsequent survey responses. For example, Bridge et al. (1977) demonstrated that individuals who participated in a survey interview about health subsequently considered the topic to be more important. And this increased importance of the topic can be translated into changed behavior. For example, people interviewed about politics are subsequently more likely to vote in elections (Granberg & Holmberg, 1992; Kraut & McConahay, 1973; Yalch, 1976). Even answering a single survey question about one's intention to vote increases the likelihood that an individual will turn out to vote on election day (Greenwald, Carnot, Beach, & Young, 1987).

Finally, panel survey respondents may want to appear consistent in their responses across waves. Therefore, people may be reluctant to report opinions or behaviors that appear inconsistent with what they had reported during earlier interviews. The desire to appear consistent could mask genuine changes over time.

Researchers can capitalize on the strengths of each of the designs discussed above by incorporating both cross-sectional and panel surveys into a single study. If, for example, a researcher is interested in conducting a two-wave panel survey but is concerned about carry-over effects, he or she could conduct an additional cross-sectional survey at the second wave. That is, the identical questionnaire could be administered to both the panel respondents and to an independent sample drawn from the same population. Significant differences between the data collected from these two samples would suggest that carry-over effects were, in fact, a problem in the panel survey. In effect, the cross-sectional survey respondents can serve as a "control group" against which panel survey respondents can be compared.

### Experiments within Surveys

Additional evidence of causal processes can be documented in surveys by building in experiments. If respondents are randomly assigned to "treatment" and "control" groups, differences between the two groups can then be attributed to the treatment. Every one of the survey designs described above can be modified to incorporate experimental manipulations. Some survey respondents (selected at random) can be exposed to one version of a questionnaire, whereas other respondents are exposed to another version. Differences in responses can then be attributed to the specific elements that were varied.

Many social psychologists are aware of examples of survey research that have incorporated experiments to explore effects of question order and question wording (see, e.g., Schuman & Presser, 1981). Less salient are the abundant examples of experiments within surveys that have been conducted to explore other social psychological phenomena.

**RACISM.** One experimental study within a survey was reported by Kinder and Sanders (1990), who were interested in the impact of public debates on public opinion on affirmative action. Sometimes, opponents of affirmative action have characterized it as entailing reverse discrimination against qualified White candidates; other times, opponents have characterized affirmative action as giving unfair advantages to minority candidates. Did this difference in framing change the way the general public formed opinions on the issue?

To answer this question, Kinder and Sanders (1990) asked White respondents in a national survey about whether they favored or opposed affirmative action programs in hiring and promotions and in college admissions. Some respondents were randomly assigned to receive a description of opposition to affirmative action as emanating from the belief that it involves reverse discrimination. Other respondents, again selected randomly, were told that opposition to affirmative action emanates from the belief that it provides unfair advantages to minorities.

This experimental manipulation of the framing of opposition did not alter the percentages of people who said they favored or opposed affirmative action, but it did alter the processes by which those opinions were formed. When affirmative action was framed as giving unfair advantage to minorities (thereby making minority group members salient), it evoked more anger, disgust, and fury from respondents, and opinions were based more on general racial prejudice, on intolerance of diversity in society, and on belief in general moral decay in society. But when affirmative action was framed as reverse discrimination against qualified Whites (thereby making Whites more salient), opinions were based more on the perceived material interests of the respondent and of Whites as a group.

Because Kinder and Sanders (1990) analyzed data from a national survey, respondents varied a great deal in terms of their political expertise. Capitalizing on this diversity, Kinder and Sanders found that the impact of framing was concentrated nearly exclusively among political novices. This reinforced the implication of Krosnick and Kinder's (1990) finding regarding political expertise in their research on news media priming described earlier.

Sniderman and Tetlock (1986) and Sniderman, Tetlock, & Peterson (1993) have also conducted experiments within surveys to assess whether conservative values encourage racial prejudice in judgments about who is entitled to public assistance and who is not. In their studies, respondents were told about a hypothetical person in need of public assistance. Different respondents were randomly assigned to receive different descriptions of the person, varying in terms of previous work history, marital and parental status, age, and race. Interestingly, conservatives did not exhibit prejudice against Blacks when deciding whether he or she should receive public assistance, even when the person was said to have violated traditional values (e.g., by being a single parent or having a history of being an unreliable worker). And in fact, when presented with an individual who had a history of being a reliable worker, conservatives were substantially more supportive of public assistance for Blacks than for Whites. However, conservatives were significantly less supportive of public policies designed to assist Blacks as a group and were more likely to believe that Blacks are irresponsible and lazy. Sniderman and Tetlock (1986) concluded that a key condition for the expression of racial discrimination is therefore a focus on groups rather than individual members of the groups and that a generally conservative orientation does not encourage individual-level discrimination.

Peffley and Hurwitz (1997; Peffley, Hurwitz, & Sniderman, 1997) also conducted experiments within surveys to explore the impact of racial stereotypes on judgments regarding crime. These investigators hypothesized that although some Americans may hold negative stereotypes of Blacks, those stereotypes may only be used to make judgments about criminal acts or public policies regarding crime when the perpetrators have characteristics consistent with those negative stereotypes. Therefore, when debates about government policy focus on counterstereotypic African American perpetrators, public attitudes may not be especially driven by stereotypes.

In one experiment, respondents in a representative sample survey were told about a man accused of committing a crime and were asked how likely he was to have actually committed it and how likely he was to commit a similar crime in the future. Respondents were randomly assigned to be told that the accused was either Black or White, and they were randomly assigned to be told that the crime was either violent or not violent. When the perpetrator was Black and the crime was violent (and thereby consistent with some negative stereotypes of Black criminals), respondents who held especially negative stereotypes of Blacks were more likely than others to say the person had committed the crime and would do so again. But when the crime was not violent or when the perpetrator was White, stereotypes of Blacks had no impact on judgments of guilt or likelihood of recidivism. In another experiment, respondents with especially negative stereotypes of Blacks were more opposed than others to furlough programs for Blacks convicted of committing violent crimes, but were not especially opposed to furlough programs for Whites or for Blacks who were described as having been model prisoners. This suggests that stereotypes can have relatively little impact on public policy debates if discussions focus on counterstereotypic perpetrators.

**MOOD AND LIFE SATISFACTION.** Schwarz and Clore (1983) conducted an experiment in a survey to explore mood and misattribution. They hypothesized that general affective states can sometimes influence judgments via misattribution. Specifically, these investigators presumed that weather conditions (sunny vs. cloudy) influence people's moods, which in turn may influence how happy they say they are with their lives. This presumably occurs because people misattribute their current mood to the general conditions of their lives, rather than to the weather conditions that happen to be occurring when they are asked to make the judgment. As a result, when people are in good moods, they may overstate their happiness with their lives.

To test this hypothesis, Schwarz and Clore (1983) conducted telephone interviews with people on either sunny or cloudy days. Among respondents who were randomly selected to be asked simply how happy they were with their lives, those interviewed on sunny days reported higher satisfaction than people interviewed on cloudy days. But among people randomly selected to be asked first, "By the way, how's the weather down there?", those interviewed on sunny days reported identical levels of life satisfaction to those interviewed on cloudy days. The question about the weather presumably led people to properly attribute some of their current mood to current weather conditions, thereby insulating subsequent life satisfaction judgments from influence.

**THE BENEFITS OF EXPERIMENTS WITHIN SURVEYS.** What is the benefit of doing these experiments in representative sample surveys? Couldn't they instead have been done just as well in laboratory settings with college undergraduates? Certainly, the answer to this latter question is yes; they could have been done as traditional social psychological experiments. But the value of doing the studies within representative sample surveys is at least three-fold. First, survey evidence documents that the phenomena are widespread enough to be observable in the general population. This bolsters the apparent value of the findings in the eyes of the many nonpsychologists who instinctively question the generalizability of laboratory findings regarding undergraduates.

Second, estimates of effect size from surveys provide a more accurate basis for assessing the significance that any social psychological process is likely to have in the course of daily life. Effects that seem large in the lab (perhaps because undergraduates are easily influenced) may actually be quite small and thereby much less socially consequential in the general population. And third, general population samples allow researchers to explore whether attributes of people that are homogeneous in the lab but vary dramatically in the general population (e.g., age, educational attainment) moderate the magnitudes of effects or the processes producing them (e.g., Kinder & Sanders, 1990).

## SAMPLING

Once a survey design has been specified, the next step in a survey investigation is selecting a sampling method (see, e.g., Henry, 1990; Kalton, 1983; Kish, 1965; Sudman, 1976). One need not look far in the social science literature to find examples where the conclusions of studies were dramatically altered when proper sampling methods were used (see, e.g., Laumann, Michael, Gagnon, & Michaels, 1994). In this section, we explain a number of sampling methods and discuss their strengths and weaknesses. In this discussion, the term *element* is used to refer to the individual unit about which information is sought. In most studies, elements are the people who make up the population of interest, but elements can also be groups of people, such as families, corporations, or departments. A *population* is the complete group of elements to which one wishes to generalize findings obtained from a sample.

### Probability Sampling

There are two general classes of sampling methods: nonprobability and probability sampling. *Nonprobability sampling* refers to selection procedures in which elements are not randomly selected from the population or some elements have unknown probabilities of being selected. *Probability sampling* refers to selection procedures in which elements are randomly selected from the sampling frame and each element has a known, nonzero chance of being selected. This does not require that all elements have an equal probability, nor does it preclude some elements from having a certain (1.00) probability of selection. However, it does require that the selection of each element must be independent of the selection of every other element.

Probability sampling affords two important advantages. First, researchers can be confident that a selected sample is representative of the larger population from which it was drawn only when a probability sampling method has been used. When elements have been selected through other procedures or when portions of the population had no chance of being included in the sample, there is no way to know whether the sample is representative of the population. Generalizations beyond the specific elements in the sample are therefore only warranted when probability sampling methods have been used.

The second advantage of probability sampling is that it permits researchers to precisely estimate the amount of variance present in a given data set that is due to sampling error. That is, researchers can calculate the degree to which random differences between the sample and the sampling frame are likely to have diminished the precision of the obtained estimates. Probability sampling also permits researchers to construct confidence intervals around their parameter estimates, which indicate the precision of the point estimates.

**SIMPLE RANDOM SAMPLING.** Simple random sampling is the most basic form of probability sampling. With this method, elements are drawn from the population at random, and all elements have the same chance of being selected. Simple random sampling can be done with or without replacement, where replacement refers to returning selected elements to the population, making them eligible to be selected again. In practice, sampling without replacement (i.e., so that each element has the potential to be selected only once) is most common.

Although conceptually a very straightforward procedure, in practice, simple random sampling is relatively difficult and costly to execute. Its main disadvantage is that it requires that all members of the population be identified so that elements can be independently and directly selected from the full population listing (the sampling frame). Once this has been accomplished, the simple random sample is drawn from the frame by applying a series of random numbers that lead to certain elements being chosen and others not. In many cases, it is impossible or impractical to enumerate every element of the population of interest, which rules out simple random sampling.

**SYSTEMATIC SAMPLING.** Systematic sampling is a slight variation of simple random sampling that is more convenient to execute (e.g., see Ahlgren, 1983). Like simple random sampling, systematic sampling requires that all elements be identified and listed. Based on the number of elements in the population and the desired sample size, a sampling interval is determined. For example, if a population contains 20,000 elements and a sample of 2,000 is desired, the appropriate sampling interval would be 10. That is, every 10th element would be selected to arrive at a sample of the desired size.

To start the sampling process, a random number between one and 10 is chosen, and the element on the list that corresponds to this number is included in the sample. This randomly selected number is then used as the starting point for choosing all other elements. Say, for example, the randomly selected starting point was 7 in a systematic sample with a sampling interval of 10. The 7th element on the list would be the first to be included in the sample, followed by the 17th element, the 27th element, and so forth.[1]

It is important to note that systematic sampling will yield a sample that is representative of the sampling frame from which it was drawn only if the elements composing the list have been arranged in a random order. When the elements are arranged in some nonrandom pattern, systematic sampling will not necessarily yield samples that are representative of the populations from which they are drawn. This potential problem is exacerbated when the elements are listed in a cyclical pattern. If the cyclical pattern of elements coincided with the sampling interval, one would draw a distinctly unrepresentative sample.

To illustrate this point, consider a researcher interested in drawing a systematic sample of men and women who had sought marital counseling within the last 5 years. Suppose he or she obtained a sampling frame consisting of a list of individuals meeting this criterion, arranged by couple: each husband's name listed first, followed by the wife's name. If the researcher's randomly chosen sampling interval was an even number, he or she would end up with a sample composed exclusively of women or exclusively of men, depending on the random start value. This problem is referred to as *periodicity*, and it can be easily avoided by randomizing the order of elements within the sampling frame before applying the selection scheme.

**STRATIFIED SAMPLING.** Stratified sampling is a slight variation of random and systematic sampling, where the sampling frame is divided into subgroups (i.e., strata), and the sampling process is executed separately on each stratum (e.g., see Ross, 1988; Stapp & Fulcher, 1983). In the example above, the sampling frame could be divided into categories (e.g., husbands and wives), and elements could be selected from each category by either a random or systematic method. Stratified sampling provides greater control over the composition of the sample, assuring the researcher of representativeness of the sample in terms of the stratification variable(s). When the stratification variable is related to the dependent variable of interest, stratified sampling reduces sampling error below what would result from simple random sampling.

Stratification that involves the use of the same sampling fraction in each stratum is referred to as proportional stratified sampling. Disproportional stratified sampling – using different sampling fractions in different strata – can also be done. This is typically done when a researcher is interested in reducing the standard error in a stratum where the standard deviation is expected to be high. By increasing the sampling fraction in that stratum, he or she can increase the number of elements allocated to the stratum. This is often done to ensure large enough subsamples for subpopulation analyses. For example, survey researchers sometimes increase the sampling fraction (often called oversampling) for minority groups in national surveys so that reliable parameter estimates can be generated for such subgroups.

Stratification requires that researchers know in advance which variables represent meaningful distinctions between elements in the population. In the example above, gender was assumed to be an important dimension, and substantive differences were expected to exist between men and woman who had sought marital counseling in the past 5 years. Otherwise, it wouldn't matter if the sample included only men or only women. As Kish (1965) pointed out, the magnitude of the advantage of stratification depends on the relation between the stratification variable and the variable(s) of substantive interest in a study; the stronger this relation, the greater the gain from using a stratified sampling strategy.

**CLUSTER SAMPLING.** When a population is dispersed over a broad geographic region, simple random sampling and systematic sampling will result in a sample that is also dispersed broadly. This presents a practical (and costly) challenge in conducting face-to-face

[1] Some have argued that the requirement of independence among sample elements eliminates systematic sampling as a probability sampling method, because once the sampling interval has been established and a random start value has been chosen, the selection of elements is no longer independent. Nevertheless, sampling statisticians and survey researchers have traditionally regarded systematic sampling as a probability sampling method, as long as the sampling frame has been arranged in a random order and the start value has been chosen through a random selection mechanism (e.g., Henry, 1990; Kalton, 1983; Kish, 1965). We have therefore included systematic sampling as a probability sampling method, notwithstanding the potential problem of nonindependence of element selection.

interviews, because it is expensive and time-consuming to transport interviewers to widely disparate locations, collecting data from only a small number of respondents in any one place.

To avoid this problem, researchers sometimes implement cluster sampling, which involves drawing a sample with elements in groups (called "clusters") rather than one-by-one (e.g., see Roberto & Scott, 1986; Tziner, 1987). Then all elements within a cluster are sampled. From the full geographic region of interest, the researcher might randomly select neighborhoods, for example, and collect data from all of the households in each selected neighborhood. In fact, face-to-face interviewing of the American adult population is typically done in clusters of 80 to 100 households within randomly selected neighborhoods, keeping the cost of national interviewing staffs at a manageable level.

Cluster sampling can also be implemented in multiple stages, with two or more sequential steps of random sampling; this is called *multistage* sampling (e.g., see Himmelfarb & Norris, 1987). To assemble a national sample for an in-person survey, for example, one might begin by randomly selecting 100 or so counties from among the more than 3,000 in the nation. Within each selected county, one could then randomly select a census tract; and from each selected tract, one could select a specific block. Then a certain number of households on each selected block could be randomly selected for inclusion in the sample. To do this, a researcher would only need a list of all counties in the United States, all of the census tracts in the selected counties, and all the blocks within the selected tracts, and only then one would one need to enumerate all of the households on the selected blocks, from which to finally draw the sample elements.

Cluster sampling can substantially reduce the time and cost of face-to-face data collection, but it also reduces accuracy by increasing sampling error. Members of a cluster are likely to share not only proximity, but a number of other attributes as well – they are likely to be more similar to one another along many dimensions than a sample of randomly selected individuals would be. Therefore, interviews with a cluster of respondents will typically yield less precise information about the full population than would the same number of interviews with randomly selected individuals.

Furthermore, clustering creates problems because it violates an assumption underlying most statistical tests: independence of observations. That is, all the people in a particular cluster are likely to be more similar to each other than they are to people in other clusters. For statistical tests to be unbiased, this sort of nonindependence needs to be statistically modeled and incorporated in any analysis, thus making the enterprise more cumbersome.

### Threats to Sample Representativeness

Ideally, these sampling processes will yield samples that are perfectly representative of the populations from which they were drawn. In practice, however, this virtually never occurs. Two of the four sources of error within the total survey error perspective can distort survey results by compromising representativeness: sampling error and nonresponse error.

**SAMPLING ERROR.** Sampling error refers to the discrepancy between the sample data and the true population values that are attributable to random differences between the sample and the sampling frame. When one uses a probability sample, estimates of sampling error can be calculated, representing the magnitude of uncertainty regarding obtained parameter estimates. Sampling error is typically expressed in terms of the standard error of an estimate, which refers to the variability of sample estimates around the true population value, assuming repeated sampling. That is, the standard error indicates the probability of observing sample estimates of varying distances from the true population value, assuming that an infinite number of samples of a particular size are drawn from the same population. Probability theory provides an equation for calculating the standard error for a single sample from a population of "infinite" size:

$$SE = \sqrt{\text{sample variance/sample size}}. \qquad (1)$$

Once the standard error has been calculated, it can be used to construct a confidence interval around a sample estimate, which is informative regarding the precision of the parameter estimate. For example, a researcher can be 95% confident that an observed sample statistic (e.g., the sample mean for some variable) falls within 1.96 standard errors of the true population parameter. A small standard error, then, suggests that the sample statistic provides a relatively precise estimate of the population parameter.

As Equation 1 shows, one determinant of sampling error is sample size – as sample size increases, sampling error decreases. This decrease is not linear, however. Moving from a small to a moderate sample size produces a substantial decrease in sampling error, but further increases in sample size produce smaller and smaller decrements in sampling error. Thus, researchers are faced with a trade-off between the considerable costs associated with increases in sample size and the relative gains that such increases afford in accuracy.

The formula in Equation 1 is correct only if the population size is infinite. If the population is finite, then a correction factor needs to be added to the formula for the standard error. Thus, the ratio of sample size to population size is another determinant of sampling error. Data collected from 500 people will include more sampling error if the sample was drawn from a population of 100,000 people than if the sample was drawn from a population of only 1,000 people. When sampling from relatively small populations (i.e., when the sample to population ratio is high), the following alternative sampling error formula should be used:

$$SE = \sqrt{\left(\frac{\text{sample variance}}{\text{sample size}}\right)\left(\frac{\text{population size} - \text{sample size}}{\text{population size}}\right)} \tag{2}$$

As a general rule of thumb, this correction only needs to be done when the sample contains over 5% of the population (Henry, 1990). However, even major differences in the ratio of the sample size to population size have only a minor impact on sampling error. For example, if a dichotomous variable has a 50/50 distribution in the population and a sample of 1,000 elements is drawn, the standard sampling error formula would lead to a confidence interval of approximately 6 percentage points in width. If the population were only 1,500 in size (i.e., two thirds of the elements were sampled), the confidence interval width would be reduced to 5 percentage points.

As Equations 1 and 2 illustrate, sampling error is also dependent on the amount of variance in the variable of interest. If there is no variance in the variable of interest, a sample of 1 is sufficient to estimate the population value with no sampling error. And as the variance increases, sampling error also increases. With a sample of 1,000, the distribution of a dichotomous variable with a 50/50 distribution in the population can be estimated with a confidence interval 6 percentage points in width. However, the distribution of a dichotomous variable with a 10/90 distribution would have a confidence interval of approximately 3.7 percentage points in width.

The standard formula for calculating sampling error, and that used by most computer statistical programs, is based on the assumption that the sample was drawn using simple random sampling. When another probability sampling method has been used, the sampling error may actually be slightly higher or slightly lower than the standard formula indicates. This impact of sampling strategy on sampling error is called a design effect. Defined more formally, the design effect associated with a probability sample is "the ratio of the actual variance of a sample to the variance of a simple random sample of the same elements" (Kish, 1965, p. 258).

Any probability sampling design that uses clustering will have a design effect in excess of 1.0. That is, the sampling error for cluster sampling will be higher than the sampling error for simple random sampling. Any stratified sampling design, on the other hand, will have a design effect less than 1.0, indicating that the sampling error is lower for stratified samples than for simple random samples. Researchers should be attentive to design effects, because taking them into account can increase the likelihood of statistical tests detecting genuinely significant effects.

**NONRESPONSE ERROR.** Even when probability sampling is done for a survey, it is unlikely that 100% of the sampled elements will be successfully contacted and will agree to provide data. Therefore, most survey samples include some elements from whom no data are gathered.[2] A survey's findings may be subject to nonresponse error to the extent that the sampled elements from whom no data are gathered differ systematically from those from whom data are gathered.

To minimize the potential for nonresponse error, researchers implement various procedures to encourage as many selected respondents as possible to participate (see, e.g., Dillman, 1978; Fowler, 1988; Lavrakas, 1993). Stated generally, the goal here is to minimize the apparent costs of responding, maximize the apparent rewards for doing so, and establish trust that those rewards will be delivered (Dillman, 1978). One concrete approach to accomplishing these goals is sending

[2] Most researcher use the term "sample" to refer both to (a) the set of elements that are sampled from the sampling frame from which data ideally will be gathered and (b) the final set of elements on which data actually are gathered. Because almost no survey has a perfect response rate, a discrepancy almost always exists between the number of elements that are sampled and the number of elements from which data are gathered. Lavrakas (1993) suggested that the term "sampling pool" be used to refer to the elements that are drawn from the sampling frame for use in sampling and that the term "sample" be preserved for that subset of the sampling pool from which data are gathered.

letters to potential respondents informing them that they have been selected to participate in a study and will soon be contacted to do so, explaining that their participation is essential for the study's success because of their expertise on the topic, suggesting reasons why participation will be enjoyable and worthwhile, assuring respondents of confidentiality, and informing them of the study's purpose and its sponsor's credibility. Researchers also make numerous attempts to contact hard-to-reach people and to convince reluctant respondents to participate and sometimes pay people for participation or give them gifts as inducements (e.g., pens, golf balls). Nonetheless, most telephone surveys have difficulty achieving response rates much higher than 60%, and most face-to-face surveys have difficulty achieving response rates much higher than 70%.

In even the best academic surveys with such response rates, there are significant biases in the demographic composition of samples. For example, Brehm (1993) showed that in the two leading, recurring academic national surveys of public opinion (the National Election Studies and the General Social Surveys), certain demographic groups have been routinely represented in misleading numbers. For example, young adults and old adults are underrepresented; males are underrepresented; people with the lowest levels of education are overrepresented; and people with the highest incomes are underrepresented. Likewise, Smith (1983) reported evidence suggesting that people who don't participate in surveys are likely to have a number of distinguishing demographic characteristics (e.g., living in big cities and working long hours).

Although a response rate may seem far from 100%, such a high rate of nonresponse does not necessarily mean that a study's implications about nondemographic variables are contaminated with error. If the constructs of interest are not correlated with the likelihood of participation, then nonresponse would not distort results. So investing large amounts of money and staff effort to increase response rates might not translate into higher data quality.

A particularly dramatic demonstration of this fact was reported recently by Visser, Krosnick, Marquette, and Curtin (1996). These researchers compared the accuracy of self-administered mail surveys and telephone surveys forecasting the outcomes of statewide elections in Ohio over a 15-year period. Although the mail surveys had response rates of about 20% and the telephone surveys had response rates of about 60%, the mail surveys predicted election outcomes much more accurately (average error = 1.6%) than the telephone surveys (average error = 5.2%). In addition, the mail surveys documented the demographic characteristics of voters more accurately than did the telephone surveys. Therefore, simply having a low response rate does not necessarily mean that a survey suffers from a large amount of nonresponse error.

Studies exploring the impact of response rates on correlational results have had mixed implications. For example, Brehm (1993) found that statistically correcting for demographic biases in sample composition had very little impact on the substantive implications of correlational analyses. However, Traugott, Groves, and Lepkowski (1987) reached a different conclusion. These investigators conducted identical telephone interviews with two equivalent samples of respondents. A higher response rate was achieved with one of the samples by mailing letters to them in advance, notifying them about the survey (this improved response rates from about 56% to about 70%). Correlations between some pairs of variables were the same in the two samples. Correlations between other pairs of variables were weakly positive in the sample with the lower response rate and much more strongly positive in the sample with the higher response rate. And correlations between still other pairs of variables were strongly positive in the sample with the lower response rate and zero in the sample with the higher response rate. Thus, substantive results can change substantially as response rates are improved.

Consequently, it is worthwhile to assess the degree to which nonresponse error is likely to have distorted data from any particular sample of interest. One approach to doing so involves making aggressive efforts to recontact a randomly selected sample of people who refused to participate in the survey and collect some data from these individuals. One would especially want to collect data on the key variables of interest in the study, but it can also be useful to collect data on those dimensions along which nonrespondents and respondents seem most likely to differ substantially (see Brehm, 1993). A researcher is then in a position to assess the magnitude of differences between people who agreed to participate in the survey and those who refused to do so.

A second strategy rests on the assumption that respondents from whom data were difficult to obtain (either because they were difficult to reach or because they initially declined to participate and were later persuaded to do so) are likely to be more similar to nonrespondents than are people from whom data were relatively easy to obtain. Researchers can compare responses of people who were immediately willing to participate with those of people who had to be recontacted and persuaded to participate. The smaller the discrepancies between these groups, the less of a threat nonresponse error would seem to be (though see Lin & Schaeffer, 1995).

**COVERAGE ERROR.** One other sort of possible error deserves mention at this point: coverage error. For reasons of economy, researchers sometimes draw probability samples not from the full set of elements in a population of interest but rather from more limited sampling frames. The greater the discrepancy between the population and the sampling frame, the greater potential there is for coverage error. Such error may invalidate inferences about the population that are made on the basis of data collected from the sample.

By way of illustration, many national surveys these days involve telephone interviewing. And although their goal is to represent the entire country's population, the sampling methods used restrict the sampling frame to households with working telephones. Although the vast majority of American adults do live in households with working telephones (about 95%; Congressional Information Service, 1990), there is a 5% gap between the population of interest and the sampling frame. To the extent that people in households without telephones are different from those in households with telephones, generalization of sample results may be inappropriate.

As compared with residents of households with working telephones, those in households without working telephones tend to earn much less money, have much less formal education, tend to be much younger, and are more often racial minorities (Thornberry & Massey, 1988). If attitudes toward government-sponsored social welfare programs to help the poor are especially positive in such households (as seems quite likely), then telephone surveys may underestimate popular support for such programs. Fortunately, however, households with and without telephones tend not to differ much on many behavioral and attitudinal measures that are unrelated to income (Groves & Kahn, 1979; Thornberry & Massey, 1988). Nonetheless, researchers should be aware of coverage error due to inadequate sampling frames and, when possible, should attempt to estimate and correct for such error.

### Nonprobability Sampling

Having considered probability sampling, we turn now to nonprobability sampling methods. Such methods have been used frequently in studies inspired by the recent surge of interest among social psychologists in the impact of culture on social and psychological processes (e.g., Kitayama & Markus, 1994; Nisbett & Cohen, 1996). In a spate of articles published in top journals, a sample of people from one country has been compared with a sample of people from another country, and differences between the samples have been attributed to the impact of the countries' cultures (e.g., Benet & Waller, 1995; Han & Shavitt, 1994; Heine & Lehman, 1995; Rhee, Uleman, Lee, & Roman, 1995). In order to convincingly make such comparisons and properly attribute differences to culture, of course, the sample drawn from each culture must be representative of it. And for this to be so, one of the probability sampling procedures described above must be used.

Alternatively, one might assume that cultural impact is so universal within a country that any arbitrary sample of people will reflect it. However, hundreds of studies of Americans have documented numerous variations between subgroups within the culture in social psychological processes, and even recent work on the impact of culture has documented variation within nations (e.g., Nisbett & Cohen, 1996). Therefore, it is difficult to have much confidence in the presumption that any given social psychological process is universal within any given culture, so probability sampling seems essential to permit a conclusion about differences between cultures based on differences between samples of them.

Instead, however, nearly all recent social psychological studies of culture have employed nonprobability sampling procedures. These are procedures where some elements in the population have a zero probability of being selected or have an unknown probability of being selected. For example, Heine and Lehman (1995) compared college students enrolled in psychology courses in a public and private university in Japan with college students enrolled in a psychology course at a public university in Canada. Rhee et al. (1995) compared students enrolled in introductory psychology courses at New York University with psychology majors at Yonsei University in Seoul, Korea. Han and Shavitt (1994) compared undergraduates at the University of Illinois with students enrolled in introductory communication or advertising classes at a major university in Seoul. And Benet and Waller (1995) compared students enrolled at two universities in Spain with Americans listed in the California Twin Registry.

In all of these studies, the researchers generalized the findings from the samples of each culture to the entire cultures they were presumed to represent. For example, after assessing the extent to which their two samples manifested self-enhancing biases, Heine and Lehman (1995) concluded that "people from cultures representative of an interdependent construal of the self," instantiated by the Japanese students, "do not self-enhance to the same extent as people from cultures characteristic of an independent self," instantiated by the Canadian students (p. 605). Yet the method of recruiting potential respondents for these studies rendered zero selection probabilities for large segments of the relevant populations. Consequently, it is impossible to know whether the obtained samples were representative of those populations, and it is impossible to estimate sampling error or to construct confidence intervals for parameter estimates. As a result, the statistical calculations used in these articles to compare the different samples were invalid, because they presumed simple random sampling.

More important, their results are open to alternative interpretations, as is illustrated by Benet and Waller's (1995) study. One of the authors' conclusions is that in contrast to Americans, "Spaniards endorse a 'radical' form of individualism" (p. 715). Justifying this conclusion, ratings of the terms "unconventional," "peculiar," and "odd" loaded in a factor analysis on the same factor as ratings of "admirable" and "high-ranking" in the Spanish sample, but not in the American sample. However, Benet and Waller's American college student sample was significantly younger and more homogeneous in terms of age than their sample of California twins (the average ages were 24 years and 37 years, respectively; the standard deviations of ages were 4 years and 16 years, respectively). Among Americans, young adults most likely value unconventionality more than older adults, so what may appear in this study to be a difference between countries attributable to culture may instead simply be an effect of age that would be apparent within both cultures.

The sampling method used most often in the studies described above is called *haphazard sampling*, because participants were selected solely on the basis of convenience (e.g., because they were enrolled in a particular course at a particular university). In some cases, notices seeking volunteers were widely publicized, and people who contacted the researchers were paid for their participation (e.g., Han & Shavitt, 1994). This is problematic because people who volunteer tend to be more interested in (and sometimes more knowledgeable about) the survey topic than the general public (see, e.g., Bogaert, 1996; Coye, 1985; Dollinger & Leong, 1993), and social psychological processes seem likely to vary with interest and expertise.

Yet another nonprobability sampling method is *purposive sampling*, which involves haphazardly selecting members of a particular subgroup within a population. This technique has been used in a number of social psychological studies to afford comparisons of what are called "known groups" (e.g., Hovland, Harvey, & Sherif, 1957; Webster & Kruglanski, 1994). For example, in order to study people strongly supporting prohibition, Hovland et al. (1957) recruited participants from the Women's Christian Temperance Union, students preparing for the ministry, and students enrolled in religious colleges. And to compare people who were high and low in need for closure, Webster and Kruglanski (1994) studied accounting majors and studio art majors, respectively.

In these studies, the groups of participants did indeed possess the expected characteristics, but they may have had other characteristics as well that may have been responsible for the studies' results. This is so because the selection procedures used typically yield unusual homogeneity within the "known groups" in at least some regards and perhaps many. For example, accounting majors may have much more training in mathematics and related thinking styles than studio art majors. Had more heterogeneous groups of people high and low in need for closure been studied by Webster and Kruglanski (1994), it is less likely that they would have sharply differed in other regards and less likely that such factors could provide alternative explanations for the results observed.

*Snowball sampling* is a variant of purposive sampling, where a few members of a subpopulation are located, and each is asked to suggest other members of the subpopulation for the researcher to contact. Judd and Johnson (1981) used this method in an investigation comparing people with extreme views on women's issues to people with moderate views. To assemble a sample of people with extreme views, these investigators initially contacted undergraduate women who were members of feminist organizations and then asked them to provide names of other women who were also likely to hold similar views on women's issues. Like cluster sampling, this sampling method also violates the assumption of independence of observations, complicating analyis.

Probably the best known form of nonprobability sampling is *quota sampling*, which involves selecting members of various subgroups of the population to assemble a sample that accurately reflects known characteristics of the population. Predetermined numbers of people in each of several categories are recruited to accomplish this. For example, one can set out to recruit a sample containing half men and half women, and one third people with less than high school education, one third high school graduates, and one third people with at least some college education. If quotas are imposed on a probability sampling procedure (e.g., telephone interviews done by random digit dialing) and if the quotas are based on accurate information about a population's composition (e.g., the U.S. Census), then the resulting sample may be more accurate than simple random sampling would be, though the gain would most likely be very small.

However, quotas are not usually imposed on probability sampling procedures but instead are imposed on haphazard samples. Therefore, this approach can give an arbitrary sample the patina of representativeness, when in fact only the distributions of the quota criteria match the population. A particularly dramatic illustration of this problem is the failure of preelection polls to predict that Truman would win his bid for the U.S. Presidency in 1948. Although interviewers conformed to quotas in selecting respondents, the resulting sample was quite unrepresentative in some regards not explicitly addressed by the quotas (Mosteller, Hyman, McCarthy, Marks, & Truman, 1949). A study by Katz (1942) illustrated how interviewers tend to over-sample residents of one-family houses, American-born people, and well-educated people when these dimensions are not explicit among the quota criteria.

Given all this, we urge researchers to recognize the inherent limitations of nonprobability sampling methods and to draw conclusions about populations or about differences between populations tentatively, if at all, when nonprobability sampling methods are used. Furthermore, we encourage researchers to attempt to assess the representativeness of samples they study by comparing their attributes with known population attributes in order to bolster confidence in generalization when appropriate and to temper such confidence when necessary. To scholars in disciplines that have come to recognize the necessity of probability sampling in order to describe populations (e.g., sociology and political science), social psychological research attempting to generalize from a college student sample to a nation looks silly and damages the apparent credibility of our enterprise.

Are we suggesting that all studies of college sophomores enrolled in introductory psychology courses are of minimal scientific value? Absolutely not. The value of the vast majority of social psychological laboratory experiments does not hinge on generalizing their results to a population. Instead, these studies test whether a particular process occurs at all, to explore its mechanisms, and to identify its moderators. Any demonstrations along these lines enhance our understanding of the human mind, even if the phenomena documented occur only among select groups of American college sophomores.

After an initial demonstration of an effect or process or tendency, subsequent research can assess its generality. Therefore, work such as Heine and Lehman's (1995) is valuable because it shows us that some findings are not limitlessly generalizable and sets the stage for research illuminating the relevant limiting conditions. We must be careful, though, about presuming that we know what these limiting conditions are without proper, direct, and compelling tests of our conjectures.

## QUESTIONNAIRE DESIGN AND MEASUREMENT ERROR

Once a sample is selected, the next step for a survey researcher is questionnaire design. When designing a questionnaire, a series of decisions must be made about each question. First, will it be open-ended or closed-ended? And for some closed-ended question tasks, should one use rating scales or ranking tasks? If one uses rating scales, how many points should be on the scales and how should they be labeled with words? Should respondents be explicitly offered "no-opinion" response options or should these be omitted? In what order should response alternatives be offered? How should question stems be worded? And finally, once all the questions are written, decisions must be made about the order in which they will be asked.

Every researcher's goal is to maximize the reliability and validity of the data he or she collects. Therefore, each of the above design decisions should presumably be made so as to maximize these two indicators of data quality. Fortunately, thousands of empirical studies provide clear and surprisingly unanimous advice on the issues listed above. Although a detailed review of this literature is far beyond the scope of this chapter (see Bradburn, et al., 1981; J. M. Converse & Presser, 1986; Krosnick & Fabrigar, in press; Schuman & Presser, 1981; Sudman, Bradburn, & Schwarz, 1996), we will provide a brief tour of the implications of these studies.

### Open versus Closed Questions

An open-ended question permits the respondent to answer in his or her own words (see, e.g., C. Smith, this volume, Ch. 12; Bartholomew, Henderson, & Marcia, this volume, Ch. 11). For example, one commonly asked open-ended question is "What is the most important problem facing the country today?" In contrast, a closed-ended question requires that the respondent select an answer from a set of choices offered explicitly by the researcher. A closed-ended version of the above question might ask "What is the most important facing the country today: inflation, unemployment, crime, the federal budget deficit, or some other problem?"

The biggest challenge in using open-ended questions is the task of coding responses. In a survey of 1,000 respondents, nearly 1,000 different answers will be given to the "most important problem" question if considered word-for-word. But in order to analyze these answers, they must be clumped into a relatively small number of categories. This requires that a coding scheme be developed for each open-ended question. Multiple people must read and code the answers into the categories, the level of agreement between the coders must be ascertained, and the procedure must be refined and repeated if agreement is too low. The time and financial costs of such a procedure, coupled with the added challenge of requiring interviewers to carefully transcribe answers, have led many researchers to favor closed-ended questions, which in essence ask respondents to directly code themselves into categories that the researcher specifies.

Unfortunately, when used in certain applications, closed-ended questions have distinct disadvantages. Most important, respondents tend to confine their answers to the choices offered, even if the researcher does not wish them to do so (Jenkins, 1935; Lindzey & Guest, 1951). Explicitly offering the option to specify a different response does little to combat this problem. If the list of choices offered by a question is incomplete, even the rank ordering of the choices that are explicitly offered can be different from what would be obtained from an open-ended question. Therefore, a closed-ended question can only be used effectively if its answer choices are comprehensive, and this can often be assured only if an open-ended version of the question is administered in a pretest using a reasonably large sample. Perhaps, then, researchers should simply include the open-ended question in the final questionnaire, because they will otherwise have to deal with the challenges of coding during pretesting. Also supportive of this conclusion is evidence that open-ended questions have higher reliabilities and validities than closed-ended questions (e.g., Hurd, 1932; Remmers, Marschat, Brown, & Chapman, 1923).

One might hesitate in implementing this advice because open-ended questions may themselves be susceptible to unique problems. For example, some researchers feared that open-ended questions would not work well for respondents who are not especially articulate, because they might have special difficulty explaining their feelings. However, this seems not to be a problem (England, 1948; Geer, 1988). Second, some researchers feared that respondents would be especially likely to answer open-ended questions by mentioning the most salient possible responses, not those that are truly most appropriate. But this, too, appears not to be the case (Schuman, Ludwig, & Krosnick, 1986). Thus, open-ended questions seem to be worth the trouble they take to ask and the complexities in analysis of their answers.

### Rating versus Ranking

Practical considerations enter into the choice between ranking and rating questions as well. Imagine that one wishes to determine whether people prefer to eat carrots or peas. Respondents could be asked this question directly (a ranking question), or they could be asked to rate their attitudes toward carrots and peas separately, and the researcher could infer which is preferred. With this research goal, asking the single ranking question seems preferable and more direct than asking the two rating questions. But rank-ordering a large set of objects takes much longer and is less enjoyed by respondents than a rating task (Elig & Frieze, 1979; Taylor & Kinnear, 1971). Furthermore, ranking might force respondents to make choices between objects toward which they feel identically, and ratings can reveal not only which object a respondent prefers but also how different his or her evaluations of the objects are.

Surprisingly, however, rankings are more effective than ratings, because ratings suffer from a significant problem: *nondifferentiation*. When rating a large set of objects on a single scale, a significantly number of respondents rate multiple objects identically as a result of survey satisficing (Krosnick, 1991b). That is, although these respondents could devote thought to the response task, retrieve relevant information from memory, and report differentiated attitudes toward the objects, they choose to shortcut this process instead. To do so, they choose what appears to be a reasonable point to rate most objects on the scale and select that point over and over (i.e., nondifferentiation), rather than thinking carefully about each object and rating different objects differently (see Krosnick, 1991b; Krosnick & Alwin, 1988). As a result, the reliability and validity of ranking data are superior to those of rating data (e.g., Miethe, 1985; Munson & McIntyre, 1979; Nathan & Alexander, 1985; Rankin & Grube, 1980; Reynolds & Jolly, 1980). So although rankings do not yield interval-level measures of the perceived distances between objects in respondents' minds and are more statistically cumbersome to analyze (see Alwin & Jackson, 1982), these measures are apparently more useful when a researcher's goal is to ascertain rank orders of objects.

### Rating Scale Formats

When designing a rating scale, one must begin by specifying the number of points on the scale. A great number of studies have compared the reliability and validity of scales of varying lengths (for a review, see Krosnick & Fabrigar, in press). For bipolar scales (e.g., running from positive to negative with neutral in the middle), reliability and validity are highest for about 7 points (e.g., Matell & Jacoby, 1971). In contrast, the reliability and validity of unipolar scales (e.g., running from no importance to very high importance) seem to be optimized for a bit shorter scales, approximately 5 points long (e.g, Wikman & Warneryd, 1990). Techniques such as magnitude scaling (e.g., Lodge, 1981), which offer scales with an infinite number of points, yield data of lower quality than more conventional rating scales and should therefore be avoided (e.g., Cooper & Clare, 1981; Miethe, 1985; Patrick, Bush, & Chen, 1973).

A good number of studies suggest that data quality is better when all scale points are labeled with words than when only some are (e.g., Krosnick & Berent, 1993). Furthermore, respondents are more satisfied when more rating scale points are verbally labeled (e.g., Dickinson & Zellinger, 1980). When selecting labels, researchers should strive to select ones that have meanings that divide up the continuum into approximately equal units (e.g., Klockars & Yamagishi, 1988). For example, "very good, good, and poor" is a combination that should be avoided, because the terms do not divide the continuum equally: the meaning of "good" is much closer to the meaning of "very good" than it is to the meaning of "poor" (Myers & Warner, 1968).

Researchers in many fields these days ask people questions offering response choices such as "agree–disagree," "true–false," or "yes–no" (see, e.g., Bearden, Netemeyer, & Mobley, 1993). Yet a great deal of research suggests that these response choices sets are problematic because of acquiescence response bias (see, e.g., Couch & Keniston, 1960; Jackson, 1979; Schuman & Presser, 1981). That is, some people are inclined to say "agree," "true," or "yes," regardless of the content of the question. Furthermore, these responses are more common among people with limited cognitive skills, for more difficult items, and for items later in a questionnaire, when respondents are presumably more fatigued (see Krosnick, 1991b). A number of studies now demonstrate how acquiescence can distort the results of substantive investigations (e.g., Jackman, 1973; Winkler, Kanouse, & Ware, 1982), and in a particularly powerful example, acquiescence undermined the scientific value of *The Authoritarian Personality's* extensive investigation of facism and antisemitism (Adorno, Frankel-Brunswick, Levinson, & Sanford, 1950). This damage occurs equally when dichotanous items offer just two choices (e.g., "agree" and "disagree") as when a rating scale is used (e.g., ranging from "strongly agree" to "strongly disagree").

It might seem that acquiescence can be controlled by measuring a construct with a large set of items, half of them making assertions opposite to the other half (called "item reversals"). This approach is designed to place acquiescers in the middle of the final dimension but will do so only if the assertions made in the reversals are equally extreme as the statements in the original items. This involves extensive pretesting and is therefore cumbersome to implement. Furthermore, it is difficult to write large sets of item reversals without using the word "not" or other such negations, and evaluating assertions that include negations is cognitively burdensome and error-laden for respondents, thus adding measurement error and increasing respondent fatigue (e.g., Eifermann, 1961; Wason, 1961). And even after all this, acquiescers presumably end up at the midpoint of the resulting measurement dimension, which is probably not where most belong on subtantive grounds anyway. That is, if these individuals were induced not to acquiesce but to answer the items thoughtfully, their final index scores would presumably be more valid than placing them at the midpoint.

Most important, answering an agree–disagree, true–false, or yes–no question always involves answering a comparable rating question in one's mind first. For example, if a man is asked to agree or disagree with the assertion "I am not a friendly person," he must first decide how friendly he is (perhaps concluding "very friendly") and then translate that conclusion into the appropriate selection in order to answer the question he was asked ("disagree" to the original item). It would be simpler and more direct to ask the person how friendly he is. In fact, every agree–disagree, true–false, or yes–no question implicitly requires the respondent to make a mental rating of an object along a continuous dimension, so asking about that dimension is simpler, more direct, and less burdensome. It is not surprising, then, that the reliability and validity of other rating scale and forced choice questions are higher than those of agree–disagree, true–false, and yes–no questions (e.g., Ebel, 1982; Mirowsky & Ross, 1991; Ruch & DeGraff, 1926; Wesman, 1946). Consequently, it seems best to avoid long batteries of questions in these latter formats and instead ask just a couple of questions using other rating scales and forced choice formats.

### The Order of Response Alternatives

The answers people give to closed-ended questions are sometimes influenced by the order in which the alternatives are offered. When categorical response choices are presented visually, as in self-administered questionnaires, people are inclined toward primacy effects, whereby they tend to select answer choices offered early in a list (e.g., Krosnick & Alwin, 1987; Sudman, Bradburn, & Schwarz, 1996). But when categorical answer choices are read aloud to people, recency effects tend to appear, whereby people are inclined to select the options offered last (e.g., McClendon, 1991). These effects are most pronounced among respondents low in cognitive skills and when questions are more cognitively demanding (Krosnick & Alwin, 1987; Payne, 1949/1950). All this is consistent with the theory of satisficing (Krosnick, 1991b), which posits that response order effects are generated by the confluence of a confirmatory bias in evaluation, cognitive fatigue, and a bias in memory favoring response choices read aloud most recently. Therefore, it seems best to minimize the difficulty of questions and to rotate the order of response choices across respondents.

### No-Opinion Filters and Attitude Strength

Concerned about the possibility that respondents may feel pressure to offer opinions on issues when they truly have no attitudes (e.g., P. E. Converse, 1964), questionnaire designers have often explicitly offered respondents the option to say they have no opinion. And indeed, many more people say they "don't know" what their opinion is when this is done than when it is not (e.g., Schuman & Presser, 1981). People tend to offer this response under conditions that seem sensible (e.g., when they lack knowledge on the issue; Donovan & Leivers, 1993), and people prefer to be given this option in questionnaires (Ehrlich, 1964). However, most "don't know" responses are due to conflicting feelings or beliefs (rather than lack of feelings or beliefs all together) and uncertainty about exactly what a question's response alternatives mean or what the question is asking (e.g., Coombs & Coombs, 1976). It is not surprising, then, that the quality of data collected is no higher when a "no opinion" option is offered than when it is not (e.g., McClendon & Alwin, 1993). That is, people who would have selected this option if offered nonetheless give meaningful opinions when it is not offered.

A better way to accomplish the goal of differentiating "real" opinions from "nonattitudes" is to measure the strength of an attitude using one or more follow-up questions. Krosnick and Petty (1995) proposed that strong attitudes can be defined as those that are resistant to change, are stable over time, and have powerful impact on cognition and action. Many empirical investigations have confirmed that attitudes vary in strength, and the respondent's presumed task when confronting a "don't know" response option is to decide whether his or her attitude is sufficiently weak as to be best described by selecting that option. But because the appropriate cutpoint along the strength dimension seems exceedingly hard to specify, it would seem preferable to ask people to describe where their attitude falls along the strength continuum.

However, there are many different aspects of attitudes related to their strength that are all somewhat independent of each other (see, e.g., Krosnick, Boninger, Chuang, Berent, & Carnot, 1993). For example, people can be asked how important the issue is to them personally or how much they have thought about it or how certain they are of their opinion or how knowledgeable they are about it (for details on measuring these and many other dimensions, see Wegener, Downing, Krosnick, & Petty, 1995). Each of these dimensions can help to differentiate attitudes that are crystallized and consequential from those that are not.

### Question Wording

The logic of questionnaire-based research requires that all respondents be confronted with the same stimulus (i.e., question), so any differences between people in their responses are due to real differences between the people. But if the meaning of a question is ambiguous, different respondents may interpret it differently and respond to it differently. Therefore, experienced survey researchers advise that questions always avoid ambiguity. They also recommend that wordings be easy for respondents to understand (thereby minimizing fatigue), and this can presumably be done by using short, simple words that are familiar to people. When complex or jargony words must be used, it is best to define them explicitly.

Another standard piece of advice from seasoned surveyers is to avoid double-barreled questions, which actually ask two questions at once. Consider the question, "Do you think that parents and teachers should teach middle school students about birth control options?" If a respondent feels that parents should do such teaching and that teachers should not, there is no comfortable way to say so, because the expected answers are simply "yes" or "no." Questions of this sort should be decomposed into ones that address the two issues separately.

Sometimes, the particular words used in a question stem can have a big impact on responses. For example, Smith (1987) found that respondents in a national survey were much less positive toward "people on welfare" than toward "the poor." But Schuman and Presser (1981) found that people reacted equivalently to the concepts of "abortion" and "ending pregnancy," despite the investigators' intuition that these concepts would elicit different responses. These investigators also found that more people say that a controversial behavior should be "not allowed" than say it should be "forbidden," despite the apparent conceptual equivalence of the two phrases. Thus, subtle aspects of question wording can sometimes make a big difference, so researchers should be careful to say exactly what they want to say when wording questions. Unfortunately, though, this literature does not yet offer general guidelines or principles about wording selection.

### Question Order

An important goal when ordering questions is to help establish a respondent's comfort and motivation to provide high-quality data. If a questionnaire begins with questions about matters that are highly sensitive or controversial or that require substantial cognitive effort to answer carefully or that seem poorly written, respondents may become uncomfortable, uninterested, or unmotivated and may therefore terminate their participation. Seasoned questionnaire designers advise beginning with items that are easy to understand and answer on noncontroversial topics.

Once a bit into a questionnaire, grouping questions by topic may be useful. That is, once a respondent starts thinking about a particular topic, it is presumably easier for him or her to continue to do so, rather than having to switch back and forth between topics, question by question. However, initial questions in a sequence can influence responses to later, related questions, for a variety of reasons (see Tourangeau & Rasinski, 1988). Therefore, within blocks of related questions, it is often useful to rotate question order across respondents so that any question order effects can be empirically gauged and statistically controlled for if necessary.

### Questions to Avoid

It is often of interest to researchers to study trends over time in attitudes or beliefs. To do so usually requires measuring a construct at repeated time points in the same group of respondents. An appealing shortcut is to ask people to attempt to recall the attitudes or beliefs they held at specific points in the past. However, a great deal of evidence suggests that people are quite poor at such recall, usually presuming that they have always believed what they believe at the moment (e.g., Bem & McConnell, 1970; Ross, 1989). Therefore, such questions vastly underestimate change and should be avoided.

Because researchers are often interested in identifying the causes of people's thoughts and actions, it is tempting to ask people directly why they thought a certain thing or behaved in a certain way. This involves asking people to introspect and describe their own cognitive processes, which was one of modern psychology's first core research methods (Hothersall, 1984). However, it became clear to researchers early in this century that it did not work all that well, and Nisbett and Wilson (1977) articulated an argument about why this is so. Evidence produced since their landmark paper has largely reinforced the conclusion that many cognitive processes occur very quickly and automatically "behind a black curtain" in people's minds, so they are unaware of them and cannot describe them. Consequently, questions asking for such descriptions seem best avoided as well.

## PRETESTING

Even the most carefully designed questionnaires sometimes include items that respondents find ambiguous or difficult to comprehend. Questionnaires may also include items that respondents understand perfectly well, but interpret differently than the researcher intended. Because of this, questionnaire pretesting is conducted to detect and repair such problems. Pretesting can also provide information about probable response rates of a survey, the cost and timeframe of the data collection, the effectiveness of the field organization, and the skill level of the data collection staff. A number of pretesting methods have been developed, each of which has advantages and disadvantages, as we review next.

### Pretesting Methods for Interviewer-Administered Questionnaires

**CONVENTIONAL PRETESTING.** In conventional face-to-face and telephone survey pretesting, interviewers conduct a small number of interviews (usually between 15 and 25) and then discuss their experiences with the researcher in a debriefing session (see, e.g., Bischoping, 1989; Nelson, 1985). They describe any problems they encountered (e.g., identifying questions that required further explanation, wording that was difficult to read or that respondents seemed to find confusing) and their impressions of the respondents' experiences in answering the questions. Researchers might also look for excessive item-nonresponse in the pretest interviews, which might suggest a question is problematic. On the basis of this information, researchers can make modifications to the survey instrument to increase the likelihood that the meaning of each item is clear to respondents and that the interviews proceed smoothly.

Conventional pretesting can provide valuable information about the survey instrument, especially when the interviewers are experienced survey data collectors. But this approach has limitations. For example, what constitutes a "problem" in the survey interview is often defined rather loosely, so there is potential for considerable variance across interviewers in terms of what is reported during debriefing sessions. Also, debriefing interviews are sometimes relatively unstructured, which might further contribute to variance in interviewers' reports. Of course, researchers can standardize their debriefing interviews, thereby reducing the idiosyncrasies in the reports from pretest interviewers. Nonetheless, interviewers' impressions of respondent reactions are unavoidably subjective and are likely to be imprecise indicators of the degree to which respondents actually had difficulty with the survey instrument.

**BEHAVIOR CODING.** A second method, called behavior coding, offers a more objective, standardized approach to pretesting. Behavior coding involves monitoring pretest interviews (either as they take place or via tape recordings of them) and noting events that occur during interactions between the interviewer and the respondent (e.g., Cannell, Miller, & Oksenberg, 1981). The coding reflects each deviation from the script (caused by the interviewer misreading the questionnaire, for example, or by the respondent asking for additional information or providing an initial response that was not sufficiently clear or complete). Questions that elicit frequent deviations from the script are presumed to require modification.

Although behavior coding provides a more systematic, objective approach than conventional pretest methods, it is also subject to limitations. Most important, behavior coding is likely to miss problems centering around misconstrued survey items, which may not elicit any deviations from the script.

**COGNITIVE INTERVIEWING.** To overcome this important weakness, researchers employ a third pretest method, borrowed from cognitive psychology. It involves administering a questionnaire to a small number of people who are asked to "think aloud," verbalizing whatever considerations come to mind as they formulate their responses (e.g., Forsyth & Lessler, 1991). This "think aloud" procedure is designed to assess the cognitive processes by which respondents answer questions, which presumably provides insight into the way each item is comprehended and the strategies used to devise answers. Interviewers might also ask respondents about particular elements of a survey question, such as interpretations of a specific word or phrase or overall impressions of what a question was designed to assess.

**COMPARING THESE PRETESTING METHODS.** These three methods of pretesting focus on different aspects of the survey data collection process, and one might expect that they would detect different types of interview problems. And indeed, empirical evidence suggests that the methods do differ in terms of the kinds of problems they detect, as well as in the reliability with which they detect these problems (i.e., the degree to which repeated pretesting of a particular questionnaire consistently detects the same problems).

Presser and Blair (1994) demonstrated that behavior coding is quite consistent in detecting apparent respondent difficulties and interviewer problems. Conventional pretesting also detects both sorts of potential problems, but less reliably. In fact, the correlation between the apparent problems diagnosed in independent conventional pretesting trials of the same questionnaire can be remarkably low. Cognitive interviews also tend to exhibit low reliability across trials, and they tend to detect respondent difficulties almost exclusively.

However, the relative reliability of the various pretesting methods is not necessarily informative about the validity of the insights gained from them. And one might even imagine that low reliability actually reflects the capacity of a particular method to continue to reveal additional, equally valid problems across pretesting iterations. But unfortunately, we know of no empirical studies evaluating or comparing the validity of the various pretesting methods. Much research along these lines is clearly needed.

### Self-Administered Questionnaire Pretesting

Pretesting is especially important when data are to be collected via self-administered questionnaires, because interviewers will not be available to clarify question meaning or probe incomplete answers. Furthermore, with self-administered questionnaires, the researcher must be as concerned about the layout of the questionnaire as with the content; that is, the format must be "user-friendly" for the respondent. A questionnaire that is easy to use can presumably reduce measurement error and may also reduce the potential for nonresponse error by providing a relatively pleasant task for the respondent.

Unfortunately, however, pretesting is also most difficult when self-administered questionnaires are used, because problems with item comprehension or response selection are less evident in self-administered questionnaires than face-to-face or telephone interviews. Some researchers rely on observations of how pretest respondents fill out a questionnaire to infer problems in the instrument – an approach analogous to behavior coding in face-to-face or telephone interviewing. But this is a less than optimal means of detecting weaknesses in the questionnaire.

A more effective way to pretest self-administered questionnaires is to conduct personal interviews with a group of survey respondents drawn from the target population. Researchers can use the "think aloud" procedure described above, asking respondents to verbalize their thoughts as they complete the questionnaire. Alternatively, respondents can be asked to complete the questionnaire just as they would during actual data collection, after which they can be interviewed about the experience. They can be asked about the clarity of the instructions, the question wording, and the response options. They can also be asked about their interpretations of the questions or their understanding of the response alternatives and about the ease or difficulty of responding to the various items.

## DATA COLLECTION

The survey research process culminates in the collection of data, and the careful execution of this final step is critical to success. Next, we discuss considerations relevant to data-collection mode (face-to-face, telephone, and self-administered) and interviewer selection, training, and supervision (for comprehensive discussions, see, e.g., Bradburn & Sudman, 1979; Dillman, 1978; Fowler & Mangione, 1990; Frey, 1989; Lavrakas, 1993).

### Mode

**FACE-TO-FACE INTERVIEWS.** Face-to-face data collection often requires a large staff of well-trained interviewers who visit respondents in their homes. But this mode of data collection is not limited to in-home interviews; face-to-face interviews can be conducted in a laboratory or other locations as well. Whatever the setting, face-to-face interviews involve the oral presentation of survey questions, sometimes with visual aids. Until recently, interviewers always recorded responses on paper copies of the questionnaire, which were later returned to the researcher.

Increasingly, however, face-to-face interviewers are being equipped with laptop computers, and the entire data-collection process is being regulated by computer programs. In computer-assisted personal interviewing (CAPI), interviewers work from a computer screen, on which the questions to be asked appear one by one in the appropriate order. Responses are typed into the computer, and subsequent questions appear instantly on the screen. This system can reduce some types of interviewer error, and it permits researchers to vary the specific questions each participant is asked based on responses to previous questions. It also makes the incorporation of experimental manipulations into a survey easy, because the manipulations can be written directly into the CAPI program. In addition, this system eliminates the need to enter responses into a computer after the interview has been completed.

**TELEPHONE INTERVIEWS.** Instead of interviewing respondents in person, researchers rely on telephone interviews as their primary mode of data collection. And whereas computerized data collection is a relatively recent development in face-to-face interviewing, most large-scale telephone survey organizations have been using such systems for the past decade. In fact, computer-assisted telephone interviewing (CATI) has become the industry standard, and several software packages are available to simplify computer programming. Like CAPI, CATI involves interviewers reading from a computer screen, on which each question appears in turn. Responses are entered immediately into the computer.

**SELF-ADMINISTERED QUESTIONNAIRES.** Often, questionnaires are mailed or dropped off to individuals at their homes, along with instructions on how to return the completed surveys. Alternatively, people can be intercepted on the street or in other public places and asked to compete a self-administered questionnaire, or such questionnaires can be distributed to large groups of individuals gathered specifically for the purpose of participating in the survey or for entirely unrelated purposes (e.g., during a class period or at an employee staff meeting). Whatever the method of distribution, this mode of data collection typically requires respondents to complete a written questionnaire and return it to the researcher.

Recently, however, even this very simple mode of data collection has benefited from advances in computer technology and availability. Paper-and-pencil self-administered questionnaires have sometimes been replaced by laptop computers, on which respondents proceed through a self-guided program that presents the questionnaire. When a response to each question is made, the next question appears on the screen, permitting respondents to work their way through the instrument at their own pace and with complete privacy. Computer assisted self-administered interviewing (CASAI), as it is known, thus affords all of the advantages of computerized face-to-face and telephone interviewing, along with many of the advantages of self-administered questionnaires. A very new development is audio CASAI, where the computer "reads aloud" questions to respondents, who listen on headphones and type their answers on computers.

### Choosing a Mode

Face-to-face interviews, telephone interviews, and self-administered questionnaires each afford certain advantages, and choosing among them requires trade-offs. This choice should be made with several factors in mind, including cost, characteristics of the population, sampling strategy, desired response rate, question format, question content, questionnaire length, length of the data-collection period, and availability of facilities.

**COST.** The first factor to be considered when selecting a mode of data collection is cost. Face-to-face interviews are generally more expensive than telephone interviews, which are usually more expensive than self-administered questionnaire surveys of comparable size.

**THE POPULATION.** A number of characteristics of the population are relevant to selecting a mode of data collection. For example, completion of a self-administered questionnaire requires a basic proficiency in reading and, depending on the response format, perhaps writing. Clearly this mode of data collection is inappropriate if a nonnegligible portion of the population being studied does not meet this minimum proficiency requirement. Some level of computer literacy is necessary if CASAI is to be used, and again, this may be an inappropriate mode of data collection if a nonnegligible portion of the population is not computer literate. Motivation is another relevant characteristic – researchers who suspect that respondents may be unmotivated to participate in the survey should select a mode of data collection that involves interaction with trained interviewers. Skilled interviewers can often increase response rates by convincing individuals of the value of the survey and persuading them to participate and provide high-quality data (Cannell, Oksenberg, & Converse, 1977; Marquis, Cannell, & Laurent, 1972).

**SAMPLING STRATEGY.** The sampling strategy to be used may sometimes suggest a particular mode of data collection. For example, some preelection polling organizations draw their samples from lists of currently registered voters. Such lists often provide only names and mailing addresses, which limits the mode of data collection to face-to-face interviews or self-administered surveys.

**DESIRED RESPONSE RATE.** Self-administered mail surveys typically achieve very low response rates, often less than 50% of the original sample when a single mailing is used. Techniques have been developed to yield strikingly high response rates for these surveys, but they are complex and more costly (see Dillman, 1978). Face-to-face and telephone interviews often achieve much higher response rates, which reduces the potential for nonresponse error.

**QUESTION FORM.** If a survey includes open-ended questions, face-to-face or telephone interviewing is often preferable, because interviewers can, in a standardized way, probe incomplete or ambiguous answers to ensure the usefulness and comparability of data across respondents.

**QUESTION CONTENT.** If the issues under investigation are sensitive, self-administered questionnaires may provide respondents with a greater sense of privacy and may therefore elicit more candid responses than telephone interviews and face-to-face interviews (e.g., Bishop & Fisher, 1995; Cheng, 1988; Wiseman, 1972).

**QUESTIONNAIRE LENGTH.** Face-to-face data collection permits the longest interviews, an hour or more. Telephone interviews are typically quite a bit shorter, usually lasting no more than 30 min, because respondents are often uncomfortable staying on the phone for longer. With self-administered questionnaires, response rates typically decline as questionnaire length increases, so they are generally kept even shorter.

**LENGTH OF DATA COLLECTION PERIOD.** Distributing questionnaires by mail requires significant amounts of time, and follow-up mailings to increase response rates further increase the overall turnaround time. Similarly, face-to-face interview surveys typically require a substantial length of time in the field. In contrast, telephone interviews can be completed in very little time, within a matter of days.

**AVAILABILITY OF STAFF AND FACILITIES.** Self-administered mail surveys require the fewest facilities and can be completed by a small staff. Face-to-face or telephone interview surveys are most easily conducted with a large staff of interviewers and supervisors. And ideally, telephone surveys are conducted from a central location with sufficient office space and telephone lines to accommodate a staff of interviewers, which need not be large.

## Interviewing

When data are collected face-to-face or via telephone, interviewers play key roles. We therefore review the role of interviewers, as well as interviewer selection, training, and supervision (see J. M. Converse & Schuman, 1974; Fowler & Mangione, 1986, 1990; Saris, 1991).

**THE ROLE OF THE INTERVIEWER.** Survey interviewers usually have three responsibilities. First, they are often responsible for locating and gaining cooperation from respondents. Second, interviewers are responsible to "train and motivate" respondents to provide thoughtful, accurate answers. And third, interviewers are responsible for executing the survey in a standardized way. The second and third responsibilities do conflict with one another. But providing explicit cues to the respondent about the requirements of the interviewing task can be done in a standardized way while still establishing rapport.

**SELECTING INTERVIEWERS.** It is best to use experienced, paid interviewers, rather than volunteers or students, because the former approach permits the researcher to be selective and choose only the most skilled and qualified individuals. Furthermore, volunteers or students often have an interest or stake in the substantive outcome of the research, and they may have expectancies that can inadvertently bias data collection.

Whether they are to be paid for their work or not, all interviewers must have good reading and writing skills, and they must speak clearly. Aside from these basic requirements, few interviewer characteristics have been reliably associated with higher data quality (Bass & Tortora, 1988; Sudman & Bradburn, 1982). However, interviewer characteristics can sometimes affect answers to questions relevant to those characteristics.

One instance where interviewer race may have had impact along these lines involved the 1989 Virginia gubernatorial race. Preelection polls showed Black candidate Douglas Wilder with a very comfortable lead over his White opponent. On election day, Wilder did win the election, but by a slim margin of 0.2%. According to Finkel, Guterbock, and Borg (1991), the overestimation of support for Wilder was due at least in part to social desirability. Some survey participants apparently believed it was socially desirable to express support for the Black candidate, especially when their interviewer was Black. Therefore, these respondents overstated their likelihood of voting for Wilder.

Likewise, Robinson and Rohde (1946) found that the more clearly identifiable an interviewer was as being Jewish, the less likely respondents were to express anti-Jewish sentiments. Schuman and Converse (1971) found more favorable views of Blacks were expressed to Black interviewers, though no race-of-interviewer effects appeared on numerous items that did not explicitly ask about liking of Blacks (see also Hyman, Feldman, & Stember, 1954). It seems impossible to eliminate the impact of interviewer race on responses, so it is preferable to randomly assign interviewers to respondents and then statistically control for interview race and the match between interviewer race and respondent race in analyses of data on race-related topics. More broadly, incorporating interviewer characteristics in statistical analyses of survey data seems well worthwhile and minimally costly.

**TRAINING INTERVIEWERS.** Interviewer training is an important predictor of data quality (Fowler &

Mangione, 1986, 1990). Careful interviewer training can presumably reduce random and systematic survey error due to interviewer mistakes and nonstandardized survey implementation across interviewers. It seems worth the effort, then, to conduct thorough, well-designed training sessions, especially when one is using inexperienced and unpaid interviewers (e.g., students as part of a class project). Training programs last 2 days or longer at some survey research organizations, because shorter training programs do not adequately prepare interviewers, resulting in substantial reductions in data quality (Fowler & Mangione, 1986, 1990).

In almost all cases, training should cover topics such as

- how to use all interviewing equipment,
- procedures for randomly selecting respondents within households,
- techniques for eliciting survey participation and avoiding refusals,
- opportunities to gain familiarity with the survey instrument and to practice administering the questionnaire,
- instructions regarding how and when to probe incomplete responses,
- instructions on how to record answers to open- and closed-ended questions, and
- guidelines for establishing rapport while maintaining a standardized interviewing atmosphere.

Training procedures can take many forms (e.g., lectures, written training materials, observation of real or simulated interviews), but it is important that at least part of the training session involve supervised practice interviewing. Pairs of trainees can take turns playing the roles of interviewer and respondent, for example. And role playing might also involve the use of various "respondent scripts" that present potential problems for the interviewer to practice handling.

**SUPERVISION.** Carefully monitoring ongoing data collection permits early detection of problems and seems likely to improve data quality. In self-administered surveys, researchers should monitor incoming data for signs that respondents are having trouble with the questionnaire. In face-to-face or telephone surveys, researchers should maintain running estimates of each interviewer's average response rate, level of productivity, and cost per completed interview.

The quality of each interviewer's completed questionnaires should be monitored, and if possible, some of the interviews themselves should be supervised. When surveys are conducted by telephone, monitoring the interviews is relatively easy and inexpensive and should be done routinely. When interviews are conducted face-to-face, interviewers can tape record some of their interviews to permit evaluation of each aspect of the interview.

**VALIDATION.** When data collection occurs from a single location (e.g., telephone interviews that are conducted from a central phone bank), researchers can be relatively certain that the data are authentic. When data collection does not occur from a central location (e.g., face-to-face interviews or telephone interviews conducted from interviewers' homes), researchers might be less certain. It may be tempting for interviewers to falsify some of the questionnaires that they turn in, and some occasionally do. To guard against this, researchers often establish a procedure for confirming that a randomly selected subset of all interviews did indeed occur (e.g., recontacting some respondents and asking them about whether the interview took place and how long it lasted).

## CONCLUSIONS

Over the last several decades, social psychological researchers have come to more fully appreciate the complexity of social thought and behavior. In domain after domain, simple "main effect" theories have been replaced by more sophisticated theories involving moderated effects. Statements such as "People behave like this" are being replaced by "Certain types of people behave like this under these circumstances." More and more, social psychologists are recognizing that psychological processes apparent in one social group may operate differently among other social groups and that personality factors, social identity, and cultural norms can have a profound impact on the nature of many social psychological phenomena.

And yet, the bulk of research in the field continues to be conducted with a very narrow, homogeneous base of participants – the infamous college sophomores. As a result, some scholars have come to question the generalizability of social psychological findings, and some disciplines look with skepticism at the bulk of our empirical evidence. Although we know a lot about the way college students behave in contrived laboratory settings, such critics argue, we know considerably less about the way all other types of people think and behave in their real-world environments.

In this chapter, we have provided an overview of a research methodology that permits social psychologists to address this concern about the meaning and value of our work. Surveys enable scholars to explore social psychological phenomena with samples that accurately represent the population about whom generalizations are to be made. And the incorporation of experimental manipulations into survey designs offers special scientific power. We believe that these advantages of survey research make it a valuable addition to the methodological arsenal available to social psychologists. The incorporation of this methodology into a full program of research enables researchers to triangulate across measures and methods, providing more compelling evidence of social psychological phenomena than any single methodological approach can.

Perhaps a first step for many scholars in this direction might be to make use of the many archived survey data sets that are suitable for secondary analysis. The University of Michigan is the home of the Inter-University Consortium for Political and Social Research (ICPSR), which stores and makes available hundreds of survey data sets on a wide range of topics, dating back at least five decades. Regardless of what topic is of interest to a social psychologist, relevant surveys are likely to have been done that could be usefully reanalyzed. And the cost of accessing these data is quite minimal to scholars working at academic institutions that are members of ICPSR and only slightly more to others. Also, the Roper Center at the University of Connecticut archives individual survey questions from thousands of surveys, some in ICPSR and some not. They do computer-based searches for questions containing particular key words or addressing particular topics, allowing social psychologists to make use of data sets collected as long ago as the 1940s.

So even when the cost of conducting an original survey is prohibitive, survey data sets have a lot to offer social psychologists. We hope that more social psychologists will take advantage of the opportunity to collect and analyze survey data in order to strengthen our collective enterprise. Doing so may require somewhat higher levels of funding than we have had in the past, but our theoretical richness, scientific credibility, and impact across disciplines are likely to grow as a result, in ways that are well worth the price.

## REFERENCES

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. (1950). *The authoritarian personality*. New York: Harper & Row.

Ahlgren, A. (1983). Sex differences in the correlates of cooperative and competitive school attitudes. *Developmental Psychology, 19*, 881–888.

Alwin, D. F., Cohen, R. L., & Newcomb, T. M. (1991). *The women of Bennington: A study of political orientations over the life span*. Madison: University of Wisconsin Press.

Alwin, D. F., & Jackson, D. J. (1982). Adult values for children: An application of factor analysis to ranked preference data. In K. F. Schuessler (Ed.), *Sociological methodology 1980*. San Fransisco: Jossey-Bass.

Babbie, E. R. (1990). *Survey research methods*. Belmont, CA: Wadsworth.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Bass, R. T., & Tortora, R. D. (1988). A comparison of centralized CATI facilities for an agricultural labor survey. In R. M. Groves, P. P. Beimer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 497–508). New York: Wiley.

Bearden, W. Q., Netemeyer, R. G., & Mobley, M. F. (1993). *Handbook of marketing scales*. Newbury Park, CA: Sage.

Bem, D. J., & McConnell, H. K. (1970.). Testing the self-perception explanation of dissonance phenomena: On the salience of premanipulation attitudes. *Journal of Personality and Social Psychology, 14*, 23–31.

Benet, V., & Waller, N. G. (1995). The big seven factor model of personality description: Evidence for its cross-cultural generality in a Spanish sample. *Journal of Personality and Social Psychology, 69*, 701–718.

Bischoping, K. (1989). An evaluation of interviewer debriefing in survey pretests. In C. F. Cannell, L. Oskenberg, F. J. Fowler, G. Kalton, & K. Bischoping (Eds.), *New techniques for pretesting survey questions* (pp. 15–29). Ann Arbor, MI: Survey Research Center.

Bishop, G. F., & Fisher, B. S. (1995). "Secret ballots" and self-reports in an exit-poll experiment. *Public Opinion Quarterly, 59*, 568–588.

Blalock, H. M. (1972). *Causal inferences in nonexperimental research*. New York: Norton.

Blalock, H. M. (1985). *Causal models in panel and experimental designs*. New York: Aldine.

Bogaert, A. F. (1996). Volunteer bias in human sexuality research: Evidence for both sexuality and personality differences in males. *Archives of Sexual Behavior, 25*, 125–140.

Bradburn, N. M., & Sudman, S. (1979). *Improving interview method and questionnaire design*. San Francisco: Jossey-Bass.

Bradburn, N. M., Sudman, S., & Associates. (1981). *Improving interview method and questionnaire design*. San Francisco: Jossey-Bass.

Brehm, J. (1993). *The phantom respondents*. Ann Arbor: University of Michigan Press.

Brehm, J., & Rahn, W. (1997). Individual-level evidence for the causes and consequences of social capital. *American Journal of Political Science, 41,* 999–1023.

Bridge, R. G., Reeder, L. G., Kanouse, D., Kinder, D. R., Nagy, V. T., & Judd, C. M. (1977). Interviewing changes attitudes – sometimes. *Public Opinion Quarterly, 41,* 56–64.

Byrne, D. (1971). *The attraction paradigm.* New York: Academic Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and divergent validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.

Cannell, C. F., Miller, P., & Oskenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 389–437). San Francisco: Jossey-Bass.

Cannell, C. F., Oskenberg, L., & Converse, J. M. (1977). *Experiments in interviewing techniques: Field experiments in health reporting.* 1971–1977. Hyattsville, MD: National Center for Health Services Research.

Caspi, A., Bem, D. J., & Elder, G. H., Jr. (1989). Continuities and consequences of interactional styles across the life course. *Journal of Personality, 57,* 375–406.

Cheng, S. (1988). Subjective quality of life in the planning and evaluation of programs. *Evaluation and Program Planning, 11,* 123–134.

Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An "experimental ethnography." *Journal of Personality and Social Psychology, 70,* 945–960.

Congressional Information Service. (1990). *American statistical index.* Bethesda, MD: Author.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire.* Beverly Hills, CA: Sage.

Converse, J. M., & Schuman, H. (1974). *Conversations at random.* New York: Wiley.

Converse, P. E. (1964). The nature of belief systems in the mass public. In D. E. Apter (Ed.), *Ideology and discontent* (pp. 206–261). New York: Free Press.

Cook, A. R., & Campbell, D. T. (1969). *Quasi-experiments: Design and analysis issues for field settings.* Skokie, IL: Rand McNally.

Coombs, C. H., & Coombs, L. C. (1976). 'Don't know': Item ambiguity or respondent uncertainty? *Public Opinion Quarterly, 40,* 497–514.

Cooper, D. R., & Clare, D. A. (1981). A magnitude estimation scale for human values. *Psychological Reports, 49,* 431–438.

Costa, P. T., McCrae, R. R., & Arenberg, D. (1983). Recent longitudinal research on personality and aging. In K. W. Schaie (Ed.), *Longitudinal studies of adult psychological development* (pp. 222–263). New York: Guilford Press.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology, 60,* 151–174.

Coye, R. W. (1985). Characteristics of participants and nonparticipants in experimental research. *Psychological Reports, 56,* 19–25.

Crano, W. D., & Brewer, M. B. (1986). *Principals and methods of social research.* Newton, MA: Allyn and Bacon.

Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. *Journal of Applied Psychology, 65,* 147–154.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method.* New York: Wiley.

Dollinger, S. J., & Leong, F. T. (1993). Volunteer bias and the five-factor model. *Journal of Psychology, 127,* 29–36.

Donovan, R. J., & Leivers, S. (1993). Using paid advertising to modify racial stereotype beliefs. *Public Opinion Quarterly, 57,* 205–218.

Ebel, R. L. (1982). Proposed solutions to two problems of test construction. *Journal of Educational Measurement, 19,* 267–278.

Ehrlich, H. J. (1964). Instrument error and the study of prejudice. *Social Forces, 43,* 197–206.

Eifermann, R. R. (1961). Negation: A linguistic variable. *Acta Psychologica, 18,* 258–273.

Elig, T. W., & Frieze, I. H. (1979). Measuring causal attributions for success and failure. *Journal of Personalty and Social Psychology, 37,* 621–634.

England, L. R. (1948). Capital punishment and open-end questions. *Public Opinion Quarterly, 12,* 412–416.

Finkel, S. E., Guterbock, T. M., & Borg, M. J. (1991). Race-of-interviewer effects in a preelection poll: Virginia 1989. *Public Opinion Quarterly, 55,* 313–330.

Forsyth, B. H., & Lessler, J. T. (1991). Cognitive laboratory methods: A taxonomy. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement error in surveys* (pp. 393–418). New York: Wiley.

Fowler, F. J. (1988). *Survey research methods* (2nd ed.). Beverly Hills, CA: Sage.

Fowler, F. J., Jr., & Mangione, T. W. (1986). *Reducing interviewer effects on health survey data.* Washington, DC: National Center for Health Statistics.

Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized survey interviewing.* Newbury Park, CA: Sage.

Frey, J. H. (1989). *Survey research by telephone* (2nd ed.). Newbury Park, CA: Sage.

Geer, J. G. (1988). What do open-ended questions measure? *Public Opinion Quarterly, 52,* 365–371.

Glenn, N. O. (1980). Values, attitudes, and beliefs. In O. G. Brim & J. Kagan (Eds.), *Constancy and change in human development* (pp. 596–640). Cambridge, MA: Harvard University Press.

Granberg, D. (1985). An anomaly in political perception. *Public Opinion Quarterly, 49,* 504–516.

Granberg, D., & Holmberg, S. (1992). The Hawthorne effect in election studies: The impact of survey participation on voting. *British Journal of Political Science, 22,* 240–247.

Greenwald, A. G., Carnot, C. G., Beach, R., & Young, B. (1987). Increasing voting behavior by asking people if they expect to vote. *Journal of Applied Psychology, 72,* 315–318.

Groves, R. M. (1989). *Survey errors and survey costs.* New York: Wiley.

Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone: A national comparison with personal interviews.* New York: Wiley.

Han, S., & Shavitt, S. (1994). Persuasion and culture: Advertising appeals in individualistic and collectivist societies. *Journal of Experimental and Social Psychology, 30,* 326–350.

Hansen, M. H., & Madow, W. G. (1953). *Survey methods and theory.* New York: Wiley.

Heine, S. J., & Lehman, D. R. (1995). Cultural variation in unrealistic optimism: Does the west feel more invulnerable than the east? *Journal of Personality and Social Psychology, 68,* 595–607.

Henry, G. T. (1990). *Practical sampling.* Newbury Park, CA: Sage.

Hewstone, M., Bond, M. H., & Wan, K. C. (1983). Social factors and social attribution: The explanation of intergroup differences in Hong Kong. *Social Cognition, 2,* 142–157.

Himmelfarb, S., & Norris, F. H. (1987). An examination of testing effects in a panel study of older persons. *Personality and Social Psychology Bulletin, 13,* 188–209.

Hothersall, D. (1984). *History of psychology.* New York: Random House.

Hovland, C. I., Harvey, O. J., & Sherif, M. (1957). Assimilation and contrast effects in reactions to communication and attitude change. *Journal of Personality and Social Psychology, 55,* 244–252.

Hurd, A. W. (1932). Comparisons of short answer and multiple choice tests covering identical subject content. *Journal of Educational Psychology, 26,* 28–30.

Hyman, H. A., Feldman, J., & Stember, C. (1954). *Interviewing in social research.* Chicago: University of Chicago Press.

Jackman, M. R. (1973, June). Education and prejudice or education and response-set? *American Sociological Review, 38,* 327–339.

Jackson, J. E. (1979). Bias in closed-ended issue questions. *Political Methodology, 6,* 393–424.

James, L. R., & Singh, B. H. (1978). An introduction to the logic, assumptions, and the basic analytic procedures of two-stage least squares. *Psychological Bulletin, 85,* 1104–1122.

Jenkins, J. G. (1935). *Psychology in business and industry.* New York: Wiley.

Judd, C. M., & Johnson, J. T. (1981). Attitudes, polarization, and diagnosticity: Exploring the effect of affect. *Journal of Personality and Social Psychology, 41,* 26–36.

Kalton, G. (1983). *Introduction to survey sampling.* Beverly Hills, CA: Sage.

Katz, D. (1942). Do interviewers bias poll results? *Public Opinion Quarterly, 6,* 248–268.

Kenny, D. A. (1979). *Correlation and causality.* New York: Wiley.

Kessler, R. C., & Greenberg, D. F. (1981). *Linear panel analysis: Models of quantitative change.* New York: Academic Press.

Kinder, D. R. (1978). Political person perception: The asymmetrical influence of sentiment and choice on perceptions of presidential candidates. *Journal of Personality and Social Psychology, 36,* 859–871.

Kinder, D. R., & Sanders, L. M. (1990). Mimicking political debate within survey questions: The case of White opinion on affirmative action for Blacks. *Social Cognition, 8,* 73–103.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Kitayama, S., & Markus, H. R. (1994). *Emotion and culture: Empirical studies of mutual influence.* Washington, DC: American Psychological Association.

Klockars, A. J., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement, 25,* 85–96.

Kraut, R. E., & McConahay, J. B. (1973). How being interviewed affects voting: An experiment. *Public Opinion Quarterly, 37,* 398–406.

Krosnick, J. A. (1988a). Attitude importance and attitude change. *Journal of Experimental Social Psychology, 24,* 240–255.

Krosnick, J. A. (1988b). The role of attitude importance in social evaluation: A study of policy preferences, presidential candidate evaluations, and voting behavior. *Journal of Personality and Social Psychology, 55,* 196–210.

Krosnick, J. A. (1991a). Americans' perceptions of presidential candidates: A test of the projection hypothesis. *Journal of Social Issues, 46,* 159–182.

Krosnick, J. A. (1991b). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5,* 213–236.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly, 51,* 201–219.

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly, 52,* 526–538.

Krosnick, J. A., & Alwin, D. F. (1989). Aging and susceptibility to attitude change. *Journal of Personality and Social Psychology, 57,* 416–425.

Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science, 37,* 941–964.

Krosnick, J. A., Boninger, D. S., Chuang, Y. C., Berent, M. K., & Carnot, C. G. (1993). Attitude strength: One construct or many related constructs? *Journal of Personality and Social Psychology, 65,* 1132–1151.

Krosnick, J. A., & Brannon, L. A. (1993). The impact of the Gulf War on the ingredients of presidential evaluations:

Multidimensional effects of political involvement. *American Political Science Review, 87,* 963–975.

Krosnick, J. A., & Fabrigar, L. R. (in press). *Designing great questionnaires: Insights from psychology.* New York: Oxford University Press.

Krosnick, J. A., & Kinder, D. R. (1990). Altering popular support for the president through priming: The Iran-Contra affair. *American Political Science Review, 84,* 497–512.

Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 1–24). Hillsdale, NJ: Erlbaum.

Laumann, E. O., Michael, R. T., Gagnon, J. H., & Michaels, S. (1994). *The social organization of sexuality: Sexual practices in the United States.* Chicago: University of Chicago Press.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Lin, I. F., & Shaeffer, N. C. (1995). Using survey participants to estimate the impact of nonparticipation. *Public Opinion Quarterly, 59,* 236–258.

Lindzey, G. E., & Guest, L. (1951). To repeat – checklists can be dangerous. *Public Opinion Quarterly, 15,* 355–358.

Lodge, M. (1981). *Magnitude scaling: Quantitative measurement of opinions.* Beverly Hills, CA: Sage.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Journal of Personality and Social Psychology, 98,* 224–253.

Marquis, K. H., Cannell, C. F., & Laurent, A. (1972). Reporting for health events in household interviews: Effects of reinforcement, question length, and reinterviews. In *Vital and health statistics* (Series 2, No. 45) (pp. 1–70). Washington, DC: U.S. Government Printing Office.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert Scale items? Study I: Reliability and validity. *Educational and Psychological Measurement, 31,* 657–674.

McClendon, M. J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods and Research, 20,* 60–103.

McClendon, M. J., & Alwin, D. F. (1993). No-opinion filters and attitude measurement reliability. *Sociological Methods and Research, 21,* 438–464.

Miethe, T. D. (1985). The validity and reliability of value measurements. *Journal of Personality, 119,* 441–453.

Mirowsky, J., & Ross, C. E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2 × 2 index. *Social Psychology Quarterly, 54,* 127–145.

Mortimer, J. T., Finch, M. D., & Kumka, D. (1982). Persistence and change in development: The multidimensional self-concept. In P. B. Baltes & O. G. Brim, Jr. (Eds.), *Life-span development and behavior* (Vol. 4, pp. 263–312). New York: Academic Press.

Mosteller, F., Hyman, H., McCarthy, P. J., Marks, E. S., & Truman, D. B. (1949). *The pre-election polls of 1948: Report to the Committee on Analysis of Pre-Election Polls and Forecasts.* New York: Social Science Research Council.

Munson, J. M., & McIntyre, S. H. (1979). Developing practical procedures for the measurement of personal values in cross-cultural marketing. *Journal of Marketing Research, 16,* 48–52.

Myers, J. H., & Warner, W. G. (1968). Semantic properties of selected evaluation adjectives. *Journal of Marketing Research, 5,* 409–412.

Nathan, B. R., & Alexander, R. A. (1985). The role of inferential accuracy in performance rating. *Academy of Management Review, 10,* 109–115.

Nelson, D. (1985). Informal testing as a means of questionnaire development. *Journal of Official Statistics, 1,* 179–188.

Nesselroade, J. R., & Baltes, P. B. (1974). Adolescent personality development and historical change: 1970–1972. *Monographs of the Society for Research in Child Development, 39* (No. 1, Serial No. 154).

Nisbett, R. E. (1993). Violence and U.S. regional culture. *American Psychologist, 48,* 441–449.

Nisbett, R. E., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the south.* Boulder, CO: Westview Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychology Review, 84,* 231–259.

Patrick, D. L., Bush, J. W., & Chen, M. M. (1973). Methods for measuring levels of well-being for a health status index. *Health Services Research, 8,* 228–245.

Payne, S. L. (1949/1950). Case study in question complexity. *Public Opinion Quarterly, 13,* 653–658.

Peffley, M., & Hurwitz, J. (1997). Public perceptions of race and crime: The role of racial stereotypes. *American Journal of Political Science, 41,* 375–401.

Peffley, M., Hurwitz, J., & Sniderman, P. M. (1997). Racial stereotypes and Whites' political views of Blacks in the context of welfare and crime. *American Journal of Political Science, 41,* 30–60.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change.* New York: Springer-Verlag.

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results? In P. V. Marsden (Ed.), *Sociological methodology, 1994* (pp. 73–104). Cambridge, MA: Blackwell.

Rahn, W. M., Krosnick, J. A., & Breuning, M. (1994). Rationalization and derivation processes in survey studies of political candidate evaluation. *American Journal of Political Science, 38,* 582–600.

Rankin, W. L., & Grube, J. W. (1980). A comparison of ranking and rating procedures for value system measurement. *European Journal of Social Psychology, 10,* 233–246.

Remmers, H. H., Marschat, L. E., Brown, A., & Chapman, I. (1923). An experimental study of the relative difficulty of true-false, multiple-choice, and incomplete-sentence types of examination questions. *Journal of Educational Psychology, 14,* 367–372.

Reynolds, T. J., & Jolly, J. P. (1980). Measuring personal values: An evaluation of alternative methods. *Journal of Marketing Research, 17,* 531–536.

Rhee, E., Uleman, J. S., Lee, H. K., & Roman, R. J. (1995). Spontaneous self-descriptions and ethnic identities in individualistic and collectivist cultures. *Journal of Personality and Social Psychology, 69,* 142–152.

Roberto, K. A., & Scott, J. P. (1986). Confronting widowhood: The influence of informal supports. *American Behavioral Scientist, 29,* 497–511.

Robinson, D., & Rohde, S. (1946). Two experiments with an anti-semitism poll. *Journal of Abnormal and Social Psychology, 41,* 136–144.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96,* 341–357.

Ross, M. W. (1988). Prevalence of classes of risk behaviors for HIV infection in a randomly selected Australian population. *Journal of Sex Research, 25,* 441–450.

Ruch, G. M., & DeGraff, M. H. (1926). Corrections for chance and "guess" vs. "do not guess" instructions in multiple-response tests. *Journal of Educational Psychology, 17,* 368–375.

Saris, W. E. (1991). *Computer-assisted interviewing.* Newbury Park, CA: Sage.

Schuman, H., & Converse, J. M. (1971). The effects of Black and White interviewers on Black responses in 1968. *Public Opinion Quarterly, 35,* 44–68.

Schuman, H., Ludwig, J., & Krosnick, J. A. (1986). The perceived threat of nuclear war, salience, and open questions. *Public Opinion Quarterly, 50,* 519–536.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys.* San Diego, CA: Academic Press.

Schuman, H., Steeh, C., & Bobo, L. (1985). *Racial attitudes in America: Trends and interpretations.* Cambridge, MA: Harvard University Press.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45,* 513–523.

Sears, D. O. (1983). The persistence of early political predispositions: The role of attitude object and life stage. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 4, pp. 79–116). Beverly Hills, CA: Sage.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51,* 515–530.

Smith, T. W. (1983). The hidden 25 percent: An analysis of nonresponse in the 1980 General Social Survey. *Public Opinion Quarterly, 47,* 386–404.

Smith, T. W. (1987). That which we call welfare by any other name would smell sweeter: An analysis of the impact of question wording on response patterns. *Public Opinion Quarterly, 51,* 75–83.

Sniderman, P. M., & Tetlock, P. E. (1986). Symbolic racism: Problems of motive attribution in political analysis. *Journal of Social Issues, 42,* 129–150.

Sniderman, P. M., Tetlock, P. E., & Peterson, R. S. (1993). Racism and liberal democracy. *Politics and the Individual, 3,* 1–28.

Stapp, J., & Fulcher, R. (1983). The employment of APA members: 1982. *American Psychologist, 38,* 1298–1320.

Sudman, S. (1976). *Applied sampling.* New York: Academic Press.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions.* San Francisco: Jossey-Bass.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco: Jossey-Bass.

Taylor, J. R., & Kinnear, T. C. (1971). Numerical comparison of alternative methods for collecting proximity judgements. *American Marketing Association Proceeding of the Fall Conference,* 547–550.

Thornberry, O. T., Jr., & Massey, J. T. (1988). Trends in United States telephone coverage across time and subgroups. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 25–50). New York: Wiley.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103,* 299–314.

Traugott, M. W., Groves, R. M., & Lepkowski, J. M. (1987). Using dual frame designs to reduce nonresponse in telephone surveys. *Public Opinion Quarterly, 51,* 522–539.

Tziner, A. (1987). Congruency issues retested using Rineman's achievement climate notion. *Journal of Social Behavior and Personality, 2,* 63–78.

Visser, P. S., Krosnick, J. A., Marquette, J., & Curtin, M. (1996). Mail surveys for election forecasting? An evaluation of the Columbus Dispatch poll. *Public Opinion Quarterly, 60,* 181–227.

Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology, 52,* 133–142.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67,* 1049–1062.

Wegener, D. T., Downing, J., Krosnick, J. A., & Petty, R. E. (1995). Measures and manipulations of strength-related properties of attitudes: Current practice and future directions. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 455–487). Hillsdale, NJ: Erlbaum.

Weisberg, H. F., Haynes, A. A., & Krosnick, J. A. (1995). Social group polarization in 1992. In H. F. Weisberg (Ed.), *Democracy's feast: Elections in America* (pp. 241–249). Chatham, NJ: Chatham House.

Weisberg, H. F., Krosnick, J. A., & Bowen, B. D. (1996). *An introduction to survey research, polling, and data analysis* (3rd ed.). Newbury Park, CA: Sage.

Wesman, A. G. (1946). The usefulness of correctly spelled words in a spelling test. *Journal of Educational Psychology, 37,* 242–246.

Wikman, A., & Warneryd, B. (1990). Measurement errors in survey questions: Explaining response variability. *Social Indicators Research, 22,* 199–212.

Winkler, J. D., Kanouse, D. E., & Ware, J. E., Jr. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology, 67,* 555–561.

Wiseman, F. (1972). Methodological bias in public opinion surveys. *Public Opinion Quarterly, 36,* 105–108.

Yalch, R. F. (1976). Pre-election interview effects on voter turnouts. *Public Opinion Quarterly, 40,* 331–336.