CHAPTER SIXTEEN

Survey Research

JON A. KROSNICK, PAUL J. LAVRAKAS, AND NURI KIM

When social psychologists see a chapter offering to tell them how to conduct survey research, some respond by saying: "I don't do surveys, so the survey research methodology literature doesn't offer me tools of value for my research program. I do experiments, because they offer me the opportunity to document causality definitively. Surveys provide merely correlational data with no real value for inferring causal influence, so that's not my thing."

Such a statement reveals a lack of understanding of what survey research methodology involves. Surveys are not inherently correlational. Instead, surveys are defined as data collections for which a researcher (1) defines a specific population of people to be described, (2) draws a systematic and representative sample of members of the population, (3) collects data from those individuals either by asking them questions or by asking them to perform other tasks, and (4) computes statistics that properly reflect the nature of the sampling process used to select the individuals.[1] In fact, experiments can and routinely are embedded in surveys. Thus, surveys are not antithetical to experiments or "merely correlational." Instead, the huge literature on survey research methods offers social psychologists the opportunity to do much of the research they wish to do even more effectively, because that literature offers insights that will improve the value of the field's experimental and nonexperimental studies. More than ever before, social psychologists stand to benefit from having a command of the survey research literature and of the technique of survey research, for a variety of specific reasons.

First, a great deal of research in social psychology is done with questionnaires, and the design of those questionnaires is typically done either by reusing a questionnaire that another investigator used previously or by designing a new questionnaire based on intuition. Yet a large and growing literature in survey research suggests how to design questionnaires so as to yield measurements with maximum reliability and validity. Therefore, social psychologists' assessments can become more precise, and, by minimizing the distorting impact of random and systematic measurement error, researchers can maximize their chances of detecting real effects and associations with minimal sample sizes.

Second, if social psychology's main goal is to produce findings that will engage the interest of college undergraduates enrolled in the courses that we teach, then continuing to collect the vast majority of our data from such students is wise. That is, if we seek mainly to fill textbooks to sell to undergraduates as we teach them about themselves, basing our studies on people like them will make the findings especially compelling. But social psychologists are increasingly called upon to provide expert testimony in court, to advise government agencies, to consult with corporations (about how to manage their workforces and how to design and sell their products and services), to advise political candidates (about how to win elections), to consult with political interest groups (about how to influence government policy making), and to offer insights on human behavior via the news media. In such settings, the findings of our research are sometimes scrutinized by people and organizations who wish to dispute our conclusions on nonscientific grounds. Consequently,

[1] Although we describe surveys as focused on sampling from a population of people, some surveys randomly sample from a list of organizations and seek informants at the organizations to interview.

it is especially important that the research methods and resulting data on which our opinions are based provide a convincing justification for generalizing our findings beyond the subpopulation of college students.

Equally important, many social psychologists seek to have our research funded by government agencies, such as the National Science Foundation and the National Institutes of Health, and by private foundations. To justify investment of substantial funds in our work, it may be important that the work provide an empirical justification for making claims about a wide range of persons at all stages in the life cycle and living in many different social settings. Presuming that our findings generalize from college students to the broader population in the absence of supportive data is unlikely to be convincing to people making tough decisions about resource allocation. Therefore, we owe it to ourselves to confirm the generalizability of our findings by at least sometimes collecting data from broader segments of the public.

One increasingly popular approach to doing so is to post an experiment on a website and to allow any interested visitor to provide data (Chapter 17 by Maniaci and Rogge in this volume discusses methods for conducting research on the Internet). This may seem attractive, because the demographics of the participants are often obviously more diverse than college students. But a researcher carrying out such a study does not know the identities of the particular people who chose to do the study, how they happened to learn about its availability on the Web, whether the same person participated in the study multiple times, and the degree to which the entire group of participating respondents match a specifiable population (e.g., all Americans).

Another popular approach is to hire respondents through services such as Amazon's Mechanical Turk. These individuals complete an experiment in exchange for a small amount of money (Maniaci & Rogge, Chapter 17 in this volume). Yet it is again not clear who these people are and the degree to which they match a specific population of adults. Most notably, because of the nature of sources like Mechanical Turk, the participants seem unlikely to include people whose lives do not encourage them to earn modest sums by performing small tasks online. Indeed, some online experiments might be completed quite often by college students, so that the apparent diversification of the respondent pool is more of an illusion than a reality.

Put simply, using such techniques for data collection transforms social psychologists from knowing that their samples of college student participants are not representative of the larger population to not knowing the extent to which their participants are representative and having no scientific basis for making generalizations. In other words, a researcher would continue to be in a position where he or she must ask outsiders to "trust me" – to grant that his or her findings are broadly generalizable, despite the absence of strong evidence to that effect.

For all of these reasons, social psychologists have an incentive to do some of our studies with representative samples of the populations we seek to describe, so that we can have a scientific basis for making claims wherein we generalize our findings to populations. This is not to say that studies of college students are without merit. They are valuable, and they should continue to be done, as should studies of haphazard samples of other types of participants, because such work yields valuable scientific insights. But that work will be even more persuasive, insightful, and constructive if social psychologists occasionally replicate and extend their (experimental and nonexperimental) studies by collecting data from representative samples of defined populations.

This recommendation is very much in line with the observation offered by Petty and Cacioppo (1996), who noted that a laboratory study of college students "examines the viability of some more general hypothesis about the relationship between two (or more) variables and ascertains what might be responsible for this relationship. Once the relationship is validated in the laboratory, its applicability to various specific situations and populations can be ascertained" (pp. 3–4). In order to do so with regard to any specific population (even the population of American college students), social psychologists must understand and employ the techniques of sampling that are the bailiwick of survey research.

Fortunately, carrying out (experimental and nonexperimental) survey research to supplement laboratory-based studies is probably much easier than many social psychologists realize. First, TESS (Time-Sharing Experiments for the Social Sciences, www.tessexperiments.org) is a platform funded by the National Science Foundation that allows any researcher to conduct experiments embedded in surveys of representative samples of American adults at no cost. Similar opportunities are offered by other organizations in other countries (e.g., http://www.centerdata.nl/en/).

Second, many survey data sets are made available to researchers at no cost, which include measures of interest to social psychologists, including the General Social Survey, the American National Election Studies, and many other surveys available through archives such as the Interuniversity Consortium for Political and Social Research (http://www.icpsr.umich.edu) and the Roper Center for Public Opinion Research (http://www.ropercenter.uconn.edu/). Analyzing data from secondary sources is increasingly of interest to psychologists (Trzesniewski, Donnellan, & Lucas, 2010). To properly analyze such data, social psychologists must understand how they were collected and must therefore understand the basics of survey methodology.

Third, funding agencies are increasingly willing to pay the costs of primary data collection from representative samples by social psychologists. Obtaining such funding is likely to improve the *apparent* scientific value of a proposed line of work in the eyes of nonpsychologists, by permitting the empirical documentation of the generalizability of findings to populations of interest, therefore enhancing the fundability of a line of investigation, rather than decreasing it. That is, by conducting some experiments with samples that are representative of a population, social psychologists can reassure skeptics that their findings do indeed generalize. To effectively propose to conduct such work, a social psychologist can recruit a survey expert to join his or her research team, but the social psychologist can also choose to learn the insights offered by the survey methodology literature. This chapter is designed for both sorts of scholars.

Specifically, the survey research literature offers guidance to social psychologists with regard to: (1) how to collect data optimally in one of four modes (face-to-face interviews, telephone interviews, paper-and-pencil questionnaires, and electronic questionnaires delivered via a computer); (2) how to draw a sample of respondents from the population for data collection; (3) how to hire, train, and supervise interviewers (when data are collected via face-to-face or telephone interviews); (4) how to design and pretest questionnaires; (5) how to manage the data collection process; and (6) how to conduct proper statistical analyses to permit generalization of findings in light of the particular sampling approach used in a study. Whether a social psychologist wishes to manage the conduct of his or her survey fully or chooses instead to hire a firm to collect the data, full understanding of the particulars of the survey method is essential. Therefore, this chapter is designed to help social psychologists understand the logic of survey research and to benefit from the insights that professionals in the field have gained about best practices.

Defined formally, *survey research* is a specific type of field study that involves the collection of data from a sample of elements drawn systematically to be representative of a well-defined, large, and geographically diverse population (e.g., all adult women living in the United States) often, though not necessarily, through the use of a questionnaire (for more lengthy discussions, see Babbie, 1990; Fowler, 2009; Frey, 1989; Lavrakas, 1993; Sapsford, 2007; Weisberg, Krosnick, & Bowen, 1996). In order to understand how to conduct such research more effectively and efficiently to accurately describe people's thinking and action, survey researchers have done extensive research on various aspects of survey methodology. This chapter reviews high points of that literature, outlining more specifically why survey research may be valuable to social psychologists, explaining the utility of various study designs, reviewing the basics of survey sampling and questionnaire design, and describing optimal procedures for data collection.

STUDY DESIGNS

A survey-based research project can employ a variety of different designs, each of which is suitable for testing hypotheses of interest to social psychologists. In this section we review several standard designs, including cross-sectional, repeated cross-sectional, panel, and mixed designs, and discuss when each is appropriate for social psychological investigation.

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 can be used by social psychologists for documenting the prevalence of particular characteristics in a population. For example, researchers studying altruism might want to begin an investigation by documenting the frequency with which people report altruistic behaviors. Or researchers studying aggression might wish to begin their work by documenting the frequency with which people report aggressive behaviors, in order to provide a compelling initial backdrop for their in-depth study of aggressiveness.

Cross-sectional surveys can also yield correlational evidence about the directions and magnitudes of associations between pairs of variables. Such correlations

do not themselves provide evidence of the causal processes that gave rise to them. But such correlations are informative about the plausibility of a causal hypothesis. That is, if variable A is thought to be a cause of variable B but the two turn out to be uncorrelated empirically, the plausibility of the causal claim is thus diminished.

Cross-sectional surveys also offer the opportunity 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, and 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). (For discussions of both mediators and moderators, see Judd, Yzerbyt, & Muller, Chapter 25 in this volume.)

For example, consider the hypothesis that a perceiver will evaluate another person in part based on the degree to which that person holds similar attitudes. That is, attitude similarity is thought to cause attraction. An initial test of this hypothesis might be afforded by gauging whether perceivers who are more in favor of strict gun control laws are more attracted to a political candidate who also favors strict gun control laws. This can be gauged by the cross-sectional correlation between perceivers' attitudes on gun control and liking of the candidate.

But such a correlation could be attributable to the influence of attitude similarity on attraction or to the influence of attraction to the candidate on attitude similarity. That is, a perceiver might like a candidate because they share the same political party affiliation, and the candidate's articulate endorsement of strict gun control attitudes might then convince the perceiver to adjust his or her own attitude on the issue to match the candidate's. If this latter process were to be true, we would expect it to be more common among people with weak attitudes toward gun control laws, whereas deriving liking of the candidate from similarity of attitudes on gun control would presumably be more common among people whose gun control attitudes are strong (see Krosnick, 1988a). Therefore, by exploring whether the strength of the association between gun control attitude similarity and candidate liking varies with the strength of the perceiver's gun control attitude, we can generate evidence consistent with one or the other or neither of these causal claims (Krosnick, 1988b).

A single, cross-sectional survey can also 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 another group consisting of those 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 they were in the first group. This use of survey data amounts to employing them to "create" a quasi-experiment.

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

One drawback of the study by Krosnick and Kinder (1990) is that splitting a national survey sample in two parts confounds time with sample attributes. That is, the sample of respondents interviewed before November 25 is likely to have had some characteristics that distinguish them from those interviewed after November 25. Specifically, the former individuals may have been home more often and/or may have been more willing to agree to be interviewed. This leaves open the possibility that the two groups of people differed from one another not only because of the revelation of the Iran-Contra affair but for other reasons as well. This confounding can be overcome by conducting multiple independent surveys, one before an event occurs and one after. That way, the survey samples will be comparable to one another, so the impact of time can be studied more clearly.

Furthermore, the conduct of multiple independent surveys (drawing representative samples from the same population) over time offer the opportunity to generate a different type of evidence consistent with a hypothesized causal relation by assessing whether changes over time in a dependent variable parallel changes in a proposed independent variable. 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 or should precede 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 of equal treatment of whites and blacks.

Repeated cross-sectional surveys can also be used to study the impact of social events that occurred between the surveys (e.g., Kam & Ramos, 2008; 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; Kellstedt, Peterson, & Ramirez, 2010).

Panel Surveys

In a panel survey, data are collected from the same people at two or more points in time. One 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; Trzesniewski, Donnellan, & Robins, 2003). Just as with a single survey, one can gauge the prevalence of a characteristic in the population and cross-sectional associations between variables. But one can also test causal hypotheses in at least two ways. First, a researcher 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 (e.g., Blalock, 1985; Kessler & Greenberg, 1981 on the methods; see Chanley, Rudolph, & Rahn, 2000; Eveland, Hayes, Shah, & Kwak, 2005; 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. He found no projection at all to have occurred, thereby suggesting that the previously documented correlations were more likely attributable to other processes (e.g., deciding how much to like a candidate based on agreement with him or her on policy issues; 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 H. W. 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 earlier: that different sorts of people were interviewed before and after the Iran-Contra revelation.

Panel surveys do have some disadvantages. First, although people are often quite willing to participate in a single cross-sectional survey, fewer are usually 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 attrition 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. However, studies of panel attrition have generally found little impact of attrition on sample representativeness and substantive results (Alderman et al., 2001; Becketti, Gould, Lillard, & Welch, 1988; Clinton, 2001; Falaris & Peters, 1998; Fitzgerald, Gottschalk, & Moffitt, 1998a, 1998b; Watson, 2003; Zabel, 1998; Zagorsky & Rhoton, 1999; Ziliak & Kniesner, 1998).

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; Voogt & Van Kempen, 2002; Yalch, 1976). Even answering a single survey question about one's intention to vote can increase the likelihood that an individual will turn out to vote on election day (Greenwald, Carnot, Beach, & Young, 1987; cf. Mann, 2005).

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 recall having reported during earlier interviews. The desire to appear consistent could mask genuine changes over time.

Combined Use of Cross-Sectional and Panel Surveys

Researchers can capitalize on the strengths of each of the aforementioned designs 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 carryover 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 carryover 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 in the preceding sections can be modified to incorporate experimental manipulations. Some survey respondents (assigned randomly) 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 of the elements of traditional laboratory experiments can be easily implemented in the context of surveys, especially online surveys (see Maniaci & Rogge, Chapter 17 in this volume). For example, many social psychological studies exposed participants to a persuasive message (either in print, orally, or via video), and then participants answered questions measuring dependent variables. Such persuasive messages can easily be presented to respondents completing online surveys. Furthermore, telephone interviewers can read the persuasive message aloud to their respondents, and face-to-face interviewers can do the same, or can present a print message on a piece of paper, or can use their laptops to display a video presentation of the message.[2]

It is also possible to set up online networks of respondents who interact with one another in a group discussion in the context of an online survey. And such a group discussion can also be implemented with fictitious other respondents whose "behavior" is controlled by a computer (Davies & Gangadharan, 2009). Thus, group interactions that might be implemented in the lab can also be implemented with a representative sample of respondents.

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., Box-Steffensmeier, Jacobson, & Grant, 2000; Couper, Traugott, & Lamias, 2001; 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 assigned 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

[2] In principle, it is preferable for interviewers to be blind to the experimental condition to which each respondent is assigned, but this may be impractical in many instances. Therefore, it may be best simply to be sure that interviewers are unaware of the hypotheses being tested in an experiment.

regarding political expertise in their research on news media priming described earlier.

Sniderman and Tetlock (1986) and Sniderman, Tetlock and 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). 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.

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 assigned 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 assigned 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 threefold. 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 non-psychologists who instinctively question the generalizability of laboratory findings regarding undergraduates.

Second, estimates of effect sizes from surveys provide more accurate bases 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.

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).

Implicit Measurement

Social psychologists are increasingly interested in implicit measurement and are able to implement implicit assessment procedures easily in laboratory settings (see Gawronski & De Houwer, Chapter 12 in this volume, for an introduction to implicit methods). Fortunately, many such procedures can be implemented in the context of surveys as well. For example, it is now routine for computers used by face-to-face interviewers and telephone interviewers

and used by respondents for Internet surveys to record the amount of time each respondent takes to answer each question. Such data have proven to be quite valuable analytically. Furthermore, procedures such as the Implicit Association Test and the Affect Misattribution Paradigm are computer-based and can therefore easily be incorporated in survey data collection done via the Internet or via laptop computers that face-to-face interviewers bring to respondents' homes (e.g., Pasek et al., 2009; Payne et al., 2010).

For example, Pasek and colleagues (2009) and Payne and colleagues (2010) implemented the Affect Misattribution Paradigm within the context of online surveys of representative national samples of American adults. These studies assessed anti-black attitudes and compared implicit assessments with explicit assessments using traditional measures of consructs, such as stereotypes and symbolic racism. Statistical analyses documented negative associations of implicit and explicit anti-black attitudes with voting for Barack Obama in the 2008 U.S. presidential election and positive associations with voting for John McCain. The impact of implicit attitudes was partly but not completely mediated by explicit attitudes. The same findings were obtained in analyses of data collected via face-to-face interviews with a representative national sample of American adults in their homes in 2008.

SAMPLING

Once a survey design has been specified, the next step is selecting a sampling method (see, e.g., Henry, 1990; Kalton, 1983; Kish, 1965; Sudman, 1976). The social science literature describes many 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.[3] A *population* is the complete group of elements to which one wishes to generalize findings obtained from a sample.

[3] In some surveys, the goal is to describe a population of objects (e.g., airplanes) and to use people as informants to describe those objects.

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 (usually the population of interest), 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 dataset that is attributable 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.

It might seem as if social psychologists need not understand how sampling processes work – perhaps they can just rely on survey professionals to design the sample and collect the data for them, and the psychologists can simply analyze the data while trusting its representativeness. But in fact, this is not true. Many probability sampling designs incorporate complexities that cause unequal probabilities of selection and clustering. These unequal probabilities and clustering must be taken into account when doing statistical analysis in order to avoid introducing bias. Furthermore, in order to assure that a representative

sample accurately describes the population of interest, survey professionals routinely compute post-stratification weights to enhance the match of the sample to the population. Such weights are needed because even when a random sample is drawn from a population, the sample of people who complete the survey often deviates in observable ways from the population. For example, national surveys of random samples of Americans routinely overrepresent well-educated people and women, because members of these groups are more willing to participate than are less educated people and men. Because the true distributions of education and sex in the population can be known from the U.S. Census, post-stratification weights can be applied to a set of survey data to adjust the proportions of people in groups defined by education and sex to match the population. If education and sex are correlated with variables or processes of interest in the survey, post-stratification in this way will alter (and presumably improve) the accuracy of the results of statistical analyses. Even social psychologists who rely on others to collect their survey data and to compute their weights should understand how the weights are computed in order to apply them properly during analysis.

Fortunately, social psychologists can benefit from very recent developments that make the computation of weights very easy. This is partly thanks to a blue-ribbon panel of sampling experts who provided advice to the American National Election Studies on how best to implement this sort of computational exercise (see Debell & Krosnick, 2009). Pasek (2012a, 2012b) has written software that is available at no cost for implementation in R that social psychologists can use relatively easily.

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 rarely done. Doing it requires that the researcher have a list of all members of the population in advance of drawing the sample, so that elements can be independently and directly selected from the full population listing (the sampling frame). 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. This can be done for a survey of, say, all of the employees of a particular company, and it can be done for surveys of the general population of some nations other than the United States, which maintain and update a list of all citizens and noncitizen residents of their countries, such as the Netherlands. But the United States does not maintain and distribute a list of all people living in the country, so it is not possible to draw a simple random sample from the population of all Americans. Nonetheless, a random sample can be drawn from this population using other, more complex methods, as we explain later.

SYSTEMATIC SAMPLING. Systematic sampling is a variation of simple random sampling that is slightly more convenient to execute (e.g., Ahlgren, 1983; Kim, Scheufele, Shanahan, & Choi, 2011; Wright, Middleton, & Yon, 2012). 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 in this example, a random number between 1 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.[4]

[4] 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

We can only be confident that systematic sampling will yield a sample that is representative of the sampling frame from which it was drawn 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 five 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 hybrid of random and systematic sampling, where the sampling frame is divided into subgroups (i.e., strata) and the sampling process is executed either separately on each stratum (e.g., Green & Gerber, 2006; Link, Battaglia, Frankel, Osborn, & Mokdad, 2007; Ross, 1988; Stapp & Fulcher, 1983) or systematically across the entire set of strata. In the example mentioned in the preceding subsection, 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). That is, the researcher can implement sampling within genders in order to assure that the ratio of husbands to wives in the sample exactly matches the ratio in the population. 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, a researcher might increase the sampling fraction (often called oversampling) for minority groups in a national survey so that reliable parameter estimates can be generated for such subgroups. It is important to bear in mind here that a representative sample is achieved as long as every member of the population has a known, nonzero probability of being selected into the sample, even if different individuals in the population have different selection probabilities. In other words, random sampling does not require that all members of the population have the same probability of being selected.

Stratification requires that researchers know in advance which variables represent meaningful distinctions between elements in the population. In the example presented earlier, gender was known to be an important dimension, and substantive differences were expected to exist between men and women who had sought marital counseling in the past five years. Of course, if gender were uncorrelated with the dependent variables, it would not 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 in reducing sampling error from using a stratified sampling strategy. This gain is manifested by greater precision of estimates and more confidence in one's conclusions.

CLUSTER SAMPLING. When a population is dispersed over a broad geographic region, simple random sampling and systematic sampling should yield a sample that is also dispersed broadly. This presents a practical (and costly) challenge in conducting face-to-face interviews, because it is expensive and time-consuming to transport interviewers to widely

have, therefore, included systematic sampling as a probability sampling method, notwithstanding the potential problem of nonindependence of element selection.

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 ("clusters") rather than one by one (e.g., Roberto & Scott, 1986; Tziner, 1987). Then all elements within a selected cluster are sampled. From the full geographic region of interest, the researcher might randomly select census tracts, for example, and try to collect data from all of the households in each selected neighborhood.

Cluster sampling has another advantage as well: It permits drawing a random sample from a population when a researcher does not have a list of all population members. For example, if a researcher wishes to conduct a survey of a representative sample of all American residents, it is possible to purchase a set of addresses selected from the U.S. Postal Service's list of all blocks in the country. A researcher could then draw a random sample of blocks and interview everyone whose primary residence is on the selected blocks. Because everyone has an equal and known probability of being selected, this approach will yield a random sample of the nation, even though it is clustered. Face-to-face interviewing of the American adult population is typically done in clusters of households within randomly selected neighborhoods, keeping the cost of maintaining and deploying 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., Himmelfarb & Norris, 1987; Li, 2008; Shen, Wang, Guo, & Guo, 2009). 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 census block or its equivalent. 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 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 would one need to enumerate all of the households on the selected blocks, from which to finally draw the sample elements. That is, the researcher need not begin with a complete listing of all members of the population.

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 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. 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.

TYPICAL SAMPLING METHODS. In practice, each mode of survey data collection has its most popular sampling method. Face-to-face surveys of geographically distributed probability samples (e.g., of all Americans) typically involve multistage cluster sampling. In recent years, this method applied to households in the United States begins by purchasing a list of addresses based on the Delivery Sequence File (DSF) assembled by the U.S. Postal Service, and drawing a sample from it. Random Digit Dial (RDD) telephone surveys typically begin with all working area codes and all working central office codes (the next three digits after the area code) for landlines and cell phones and attach four randomly generated digits to yield a random sample of phone numbers. For mail surveys of general population samples, researchers can also begin with a list of addresses, perhaps one that is based on the U.S. Postal Service's DSF, and can draw a random sample from it.

With the spread of Internet access around the world, survey research firms have shifted a great deal of their data collection for academic and industry clients to Internet surveys. And it is possible to conduct such surveys with probability samples recruited either face-to-face, by telephone, or via mailed invitations. This notion was pioneered by Willem Saris (1998) in the Netherlands, who placed computers and telephone modems in the homes of a random sample of Dutch residents. They completed survey questionnaires regularly via their computers and modems. Saris drew his sample from the Dutch government's list of all residents of the country.

This idea was transported to the United States by a firm called Knowledge Networks (now called GfK), who recruited panels of people to complete surveys regularly via the Internet. Recruitment was originally done by Random Digit Dialing telephone calls, and in more recent years, some recruitment has been

done via mailed paper-and-pencil invitations. Computer equipment and Internet access have been provided to all participating individuals who lacked either. This panel has produced remarkably accurate measurements (Chang & Krosnick, 2009; Yeager et al., 2011).

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. Sampling error, nonresponse error, and coverage error can distort survey results by compromising representativeness, and social psychologists should understand how this can occur, so as to understand how it can be taken into account during data analysis.

SAMPLING ERROR. Sampling error refers to the discrepancies between values computed from the sample data (e.g., sample means) and the true population values. Such discrepancies are attributable to random differences between the initially chosen sample and the sampling frame from which the sample is drawn. When one uses a probability sample, estimates of the amount of sampling error can be calculated, representing the magnitude of uncertainty regarding obtained parameter estimates resulting from the fact that only a sample from the population was interviewed. 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 simultaneously from the same population. Probability theory provides an equation for calculating the standard error for a single sample from a population of "infinite" size:

$$SE = \sqrt{\text{sample variance}/\text{sample size}}. \qquad (16.1)$$

With a probability sample, 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 the true population parameter value (e.g., the population's mean value on some variable) falls in the interval that is within 1.96 standard errors of the observed statistic generated from a large sample. A small standard error, then, suggests that the sample statistic provides a relatively precise estimate of the population parameter.

As Equation 16.1 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 (e.g., 100) to a moderate sample size (e.g., 500) 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 small relative gains such increases often afford in accuracy.

The formula in Equation 16.1 is correct only if the population size is infinite. When the population is finite, a correction factor may need 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:

$$SE = \sqrt{\left(\frac{\text{sample variance}}{\text{sample size}}\right)\left(\frac{\text{population size} - \text{sample size}}{\text{population size}}\right)} \qquad (16.2)$$

As a general rule of thumb, this correction only needs to be done when the sample contains more than 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 16.1 and 16.2 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 one 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, 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* (deff). 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. The degree to which the design effect is less than 1.0 depends on the degree to which the stratification variable is related to the outcome variable. Social psychologists 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, almost all survey samples include some elements from whom no data were gathered.[5] A survey's findings may be subject to nonresponse error to the extent that the sampled elements from whom no data were gathered differ systematically and in nonnegligible ways from those from whom data were gathered.

To minimize the potential for nonresponse error, researchers have traditionally implemented various procedures to encourage as many selected respondents as possible to participate (e.g., Dillman, 1978; Fowler, 1988; Lavrakas, 2010). 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 "advance" 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., movie passes, pens, golf balls). Such material incentives are effective at increasing participation rates, especially when they are provided at the time the participation invitation is offered, rather than if they are promised to be provided after the interview is completed (e.g., Cantor, O'Hare, & O'Connor, 2008; Singer, Van Howeyk, & Maher, 2000; see Singer & Ye, 2013 for a review on incentives in surveys).

In even the best surveys with the best response rates, there are usually 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 were routinely represented in misleading numbers. For example, young adults and old adults are underrepresented, males are underrepresented, people with the highest levels of education are overrepresented, and people with the highest incomes are underrepresented. Likewise, Smith (1983) reported evidence suggesting that people who

[5] Most researchers 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.

do not participate in surveys are likely to have a number of distinguishing demographic characteristics (e.g., living in big cities and working long hours). Holbrook, Krosnick, and Pfent (2008) reported similar evidence.

In most cases, the farther a survey's response rate falls below 100%, the more a researcher can justify concern about the representativeness of the participating sample of respondents. That is, in general, as a survey's response rate drops, the risk of the so-called nonresponse error rises. In other words, the participating sample may be systematically different from the population if non-respondents are not a random subset of the population. This point of view was expressed especially clearly in a relatively recent revision of guidelines issued by the U.S. Office of Management and Budget regarding procedures for conducted federal surveys in America (Office of Information and Regulatory Affairs, 2006).

However, a high rate of nonresponse does not necessarily mean that a study's measurements of non-demographic variables are fraught with error (cf. Groves, 2006). If the constructs of interest are not correlated with the likelihood of participation, then non-response 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 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 did (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.

Other studies exploring the impact of response rates have also supported the same conclusion. 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. Holbrook et al. (2008) meta-analyzed a large set of telephone surveys with widely varying response rates and found that the accuracy of the samples in describing the population declined only very slightly as the response rate fell. Curtin, Presser, and Singer (2000) reanalyzed a survey dataset to see how much the substantive conclusions of the research differed if the researchers discarded more and more data to simulate determine what the survey's results would have been if the response rate had been lower and lower because the researchers terminated interviews earlier and earlier in the survey's field period. The results changed remarkably little.

In another study, Keeter et al. (2000) conducted two simultaneous surveys using the same questionnaire, one employing procedures to increase the response rate as much as possible, and the other taking few such steps. As expected, the former survey yielded a notably higher response rate than did the latter. But the substantive results of the two surveys differed little from one another. Lastly, Merkle and Edelman (2002) analyzed data from exit polls conducted on election day, wherein the response rate for interviewing varied widely from precinct to precinct. The response rate was essentially uncorrelated with the accuracy of the survey's measurement of voting in each precinct. Thus, an accumulating number of publications investigating a wide range of measures show that as long as a random sample is scientifically drawn from the population and thorough, professional efforts are made to collect data from all selected potential respondents, a substantial increase in a survey's response rate is not associated with a notable increase in the accuracy of the survey's results (but see Traugott, Groves, & Lepkowski, 1987).

Nonetheless, it is worthwhile to assess the degree to which nonresponse error is likely to have biased 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 (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 non-respondents 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 (but see Lin & Schaeffer, 1995 and Mazza & Enders, Chapter 24 in this volume).

COVERAGE ERROR. One other possible error deserves mention: 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 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 individuals with cell phones or living in households with landline telephones. Although the vast majority of American adults do have cell phones or live in households with working telephones, about 5% of the nation does not at any one time. To the extent that people who cannot be reached by phones are different from the rest of the population, generalization of sample results may be inappropriate.

More strikingly, an increasing number of telephone surveys today are so-called robo-polls, meaning that no human interviewers are involved. Instead, an audio recording of a survey's introduction and questions is made, and computers automatically dial randomly generated telephone numbers and play the recording to prospective respondents, who answer questions by either pushing buttons on touch-tone phones or answering orally, and voice-recognition software is used to interpret and record responses. However, because it is illegal in the United States for computers to automatically dial cell phones, robo-polls involve calls only to landline phones, which causes omission from the survey sample of the substantial portion of Americans who do not have a working landline in their homes (about 40% of adults in the United States in 2012; see Blumberg & Luke, 2012). This constitutes a substantial amount of noncoverage, and individuals without landlines are systematically different from those with landlines. Although some robo-polls have yielded reasonable accuracy in anticipating the results of elections, that accuracy appears likely to be an illusory result of adjusting results to match those of previously released high-quality surveys (Clinton & Rogers, 2012). Thus, the noncoverage bias may have been quite consequential.

Nonprobability Sampling

Social psychologists interested in making statements about the general population must employ probability sampling in order to have a scientific justification for generalization. But most social psychological studies have instead been conducted using nonprobability samples, even in domains that conceptually call for probability samples. For example, nonprobability sampling has been used frequently in studies inspired by the 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 was compared with a sample of people from another country, and differences between the samples were attributed to the impact of the countries' cultures (e.g., Benet & Waller, 1995; Hamamura, Meijer, Heine, Kamaya, & Hori, 2009; Han & Shavitt, 1994; Heine & Lehman, 1995; Kitayama, Park, Sevincer, Karasawa, & Uskul, 2009; Rhee, Uleman, Lee, & Roman, 1995). However, 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 earlier must be used in each country being compared.

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., Graham, Haidt, & Nosek, 2009; 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 reliable conclusion about differences between cultures based on differences between samples of them.

In this light, it is striking that 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 from a frame that covered the entire population of interest. Although the researchers using methods like this might argue that the comparisons are valid because the sampling frame was equivalent in the two countries (e.g., college students taking particular courses), no systematic random sampling was actually done from any representative frame, so generalization beyond the research participants is not justified.

More importantly, 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" (Benet & Waller, 1995, 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 Spaniards (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 do, 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 nonprobability sampling method used most often in the studies described earlier is called *haphazard sampling*, because respondents 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 (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 respondents 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 respondents did indeed possess the expected characteristics, but they may as well have had other characteristics 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 more training in mathematics and related thinking styles than studio art majors do. 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 rare 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 analysis. Recent developments with a related technique called "respondent-driven sampling" have sought to systematize the application of snowball sampling (Heckathorn, 1997, 2002; Salganik & Heckathorn, 2004).

Probably the best-known form of nonprobability sampling is *quota sampling*, which involves selecting members of various subgroups of the population to build a sample that accurately reflects certain 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 half of which is comprised of men and another half of women, and one-third of people with less than high school education, one-third of people with only a high school degree, and one-third of 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, although the gain most likely would 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 pre-election polls to predict that Truman would win his bid for the U.S. presidency in 1948. Although interviewers conformed to certain demographic 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 oversample residents of one-family houses, American-born people, and well-educated people when these dimensions are not explicit among the quota criteria.

Although surveys done with probability samples and collecting data via the Internet yield remarkably accurate results (Chang & Krosnick, 2009; Yeager et al., 2011), the vast majority of Internet survey data being collected around the world is based on nonprobability "volunteer" (so called opt-in) samples of respondents. This is an especially surprising development from a scientific standpoint, because survey professionals learned their lesson decades ago about the dangers of nonprobability sampling. Not only does this approach lack theoretical foundation, but it yields findings that are consistently less accurate than results produced with probability samples (e.g., Chang & Krosnick, 2009; Yeager et al., 2011). Despite claims to the contrary made by most of the companies that collect and sell such nonprobability sample Internet data, their methods appear not to be as accurate as those produced by probability samples.

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.

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, 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, however, 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 aforementioned 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 in the preceding paragraph. Although a detailed review of this literature is beyond the scope of this chapter (for reviews, see Bradburn, et al., 1981; Converse & Presser, 1986; Krosnick & Fabrigar, forthcoming; Saris & Gallhofer, 2007; Schuman & Presser, 1981; Sudman, Bradburn, & Schwarz, 1996), we provide a brief tour of the implications of these studies. John and Benet-Martinez (Chapter 18 in this volume) discuss reliability in more detail; Brewer and Crano (Chapter 2 in this volume) discuss validity.

Open vs. Closed Questions

An open-ended question permits the respondent to answer in his or her own words. For example, in political surveys one commonly asked nominal 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 verbatim answers will be given to the "most important problem" question if considered word for word. But in order to analyze these answers statistically, they must be clumped into a relatively small number of categories. This requires that a set of mutually exclusive and exhaustive codes be developed for each open-ended question. Multiple people should 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 importantly, 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; Presser, 1990). 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., Haddock & Zanna,1998; Hurd, 1932; Remmers, Marschat, Brown, & Chapman, 1923; Schuman, 2008; see also Smyth, Dillman, Christian, & Mcbride, 2009 on open-ended questions in Web surveys).

One might hesitate in implementing this advice because nominal 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 describing their thoughts, opinions, and feelings. However, this seems not to be a problem (England, 1948; Haddock & Zanna, 1998; 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 measure nominal constructs.

Another type of open-ended question seeks a number, such as the number of times that a respondent went out to the movies during the last month. Such a question can also be asked in a closed-ended format, offering ranges. For example, respondents can be asked whether they never went out to the movies, went out once or twice, or went out three or more times. Offering ranges like this might seem to simply the respondent's task by allowing him or her to answer approximately rather than exactly. But in fact, answering such a question accurately first requires the respondent to answer the open-ended version of the question in his or her own mind and then match that answer to one of the offered ranges. Thus, it would be simpler for respondents to skip the matching step and simply report the answer to the open-ended question. And the particular ranges offered by researchers are usually relatively arbitrarily chosen, yet they can manipulate respondents' answers (Courneya, Jones, Rhodes, & Blanchard, 2003; Hurd, 1999; Richardson, 2004; Schwarz, Hippler, Deutsch, & Strack, 1985; Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991). For example, if a question asks respondents how many hours he/she typically watches television during a week and offers a series of answer choices (e.g., "less than 2 hours, 3–5 hours, 6–8 hours, 9–11 hours, 12 or more hours), respondents infer that the researcher expects to obtain a normal distribution of responses, with the mode in the middle of the range, and infer that the midpoint is the most common behavior in the population. Therefore, respondents gravitate toward the middle of the offer ranges, no matter what ranges are offered. Thus, numeric questions are best asked seeking an exact number from respondents.

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 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, partly because ratings suffer from a significant problem: *nondifferentiation*. When rating a large set of objects on a single scale, a significant 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 often 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 (Krosnick, 1991b; Krosnick & Alwin, 1988). As a result, the reliability and validity of ranking data are superior to those of rating data (e.g., Harzing et al., 2009; 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 (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. Many studies have compared the reliability and validity of scales of varying lengths (for a review, see Krosnick & Fabrigar, forthcoming). For bipolar scales (e.g., running from positive to negative with neutral in the middle), reliability and validity are highest for about seven points (e.g., Matell & Jacoby, 1971; see also Alwin & Krosnick, 1991; Lozano, Garcia-Cueto, & Muñiz, 2008; Preston & Colman, 2000). 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 five 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 do more conventional rating scales and should therefore be avoided (e.g., Cooper & Clare, 1981; Miethe, 1985; Patrick, Bush, & Chen, 1973; see also Cook, Heath, & Thompson, 2001 and Couper, Tourangeau, Conrad, & Singer, 2006 on Web-based visual analogue rating scales).

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; Weng, 2004; Weijters, Cabooter, & Schillewaert, 2010). Furthermore, respondents are more satisfied when more rating scale points are verbally labeled (e.g., Dickinson & Zellinger, 1980). Researchers should strive to select labels 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" (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 (e.g., Couch & Keniston, 1960; Jackson, 1979; Schuman & Presser, 1981; Schuman & Scott, 1989). 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 (Krosnick, 1991b). A number of studies demonstrate how acquiescence can distort the results of substantive investigations (e.g., Jackman, 1973; Saris, Revilla, Krosnick, & Shaeffer, 2010; Winkler, Kanouse, & Ware, 1982), and in a particularly powerful historical example, acquiescence undermined the scientific value of *The Authoritarian Personality*'s extensive investigation of fascism and anti-Semitism (Adorno, Frankel-Brunswick, Levinson, & Sanford, 1950). This damage occurs equally when dichotomous 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 acquiescing responders 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, acquiescing respondents presumably end up at the midpoint of the resulting measurement dimension, which is probably not where most belong on substantive 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 first answering a comparable rating question in one's mind. 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 two or three questions using other rating scales and forced choice formats (e.g., Robins, Hendin, & Trzesniewski, 2001).

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., Galesic, Tourangeau, Couper, & Conrad, 2008; Krosnick & Alwin, 1987; Miller & Krosnick, 1998; Sudman et al., 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., Holbrook, Krosnick, Moore, & Tourangeau, 2007; McClendon, 1991). And when rating scales are presented visually and orally, primacy effects routinely appear. These effects are most pronounced among respondents low in cognitive skills and when questions are more cognitively demanding (Holbrook et al., 2007; Krosnick & Alwin, 1987; Payne, 1949/1950; Schuman & Presser, 1996). 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; Schuman & Scott, 1989). People tend to offer this response under conditions that seem sensible (e.g., when they lack knowledge on the issue; Donovan & Leivers, 1993; Luskin & Bullock, 2011), and people prefer to be given this option in questionnaires (Ehrlich, 1964). However, most "don't know" responses stem from 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; see also Berinsky, 1999). 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., Krosnick, Holbrook, Berent, Carson, Hanemann, Kopp, Mitchell, Presser, Ruud, Smith, Moody, Green, & Conaway, 2002; 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 cut point 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 (e.g., Krosnick, Boninger, Chuang, Berent, & Carnot, 1993; Wojcieszak, 2012). 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 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 stem from 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 surveyors is to avoid so-called 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, however, 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 engaging, noncontroversial topics.

Once into a questionnaire a bit, 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 (McFarland, 1981; Moore, 2002; Sudman et al., 1996; Tourangeau & Rasinski, 1988; Tourangeau, Rips, & Rasinski, 2000; Van de Walle & Van Ryzin, 2011; Wilson, 2010). For example, when national survey respondents were asked whether it should be possible for a married woman to obtain a legal abortion if she does not want any more children, fewer respondents expressed support after first being asked whether abortion should be legal if there is a strong chance of a serious defect in the baby (Schuman & Presser, 1981). This might be owing to a perceptual contrast effect, given that the latter seems like a more compelling justification than the former is. Also, being asked whether communist news reporters should be allowed to work in the United States in the 1940s, American survey respondents were far more likely answer affirmatively if they had previously been asked if American news reporters should be allowed to work in the Soviet Union (Schuman & Presser, 1981). This may be the result of activation of the "norm of reciprocity" at the time the second question is asked: a norm that states all parties in a dispute should be treated equally. Therefore, within blocks of related questions on a single topic, it might be 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 unless the researcher wishes to measure people's perceptions of change per se.

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 in the 20th century that it did not work well, and Nisbett and Wilson (1977) articulated an argument about why this is so. Evidence produced since their landmark work 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, unless researchers wish to measure people's perceptions of the causes of their thinking and action per se.

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 time frame 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

Pretesting questionnaires is routine in survey research, but may be done more rarely by social psychologists. Often, questionnaires designed based on intuition or tradition are deployed without determining whether they are effective measuring tools. If social psychologists wish to learn from survey researchers about how to evaluate their questionnaires before deploying them, a variety of techniques are available, as we outline next.

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 (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 non-response in the pretest interviews, which might suggest a question is problematic. On the basis of this information, researchers can modify 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 video recordings) and noting events that occur during interactions between the interviewer and the respondent (e.g., Cannell, Miller, & Oksenberg, 1981; Hess, Singer, & Bushery, 1999). 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., Beatty & Willis, 2007; Forsyth & Lessler, 1991; Willis, 2005). 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 (for a review, see Presser, Rothgeb, Couper, Lessler, Martin, Martin, & Singer, 2004).

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. Achieving this goal is a particular challenge when doing surveys via the Internet, because different browsers display text and graphics differently to different respondents (Couper, 2008). 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 face-to-face interviews with a group of survey respondents drawn from the target population. Researchers can use the previously described "think aloud" procedure, 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 "field period," during which the data are collected, 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, 2007; Fowler & Mangione, 1990; Frey, 1989; Lavrakas, 1993).

Mode

FACE-TO-FACE INTERVIEWS. National 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. For many years, interviewers recorded responses on paper copies of the questionnaire, but now face-to-face interviewers are equipped with laptop or tablet computers, and the entire data collection process is being regulated by computer software.

In computer-assisted personal interviewing (CAPI; see United Nations Economic and Social Commission for Asia and the Pacific, 1999a), 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 respondent 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 incorporated 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 sometimes rely on telephone interviewing as their primary mode of data collection, and such interviewing is almost always driven by software presenting questions on computer screens to interviewers. Responses are typed immediately into the computer. So-called computer-assisted telephone interviewing (CATI; United Nations Economic and Social Commission for Asia and the Pacific, 1999b) is the industry standard, and several software packages are available to simplify computer programming.

SELF-ADMINISTERED QUESTIONNAIRES. Self-administration is employed when paper 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.

Although paper questionnaire self-administration continues today in prominent contexts (e.g., the exit polls conducted on election days by major news media organizations), computer self-administration is now much more common. Not only are computers used for Internet surveys, but they can be used in face-to-face interviewing and telephone as well. In-home interviewers usually bring laptop or tablet computers with them and can pass those computers to their respondents, who can listen to questions being read aloud to them on headphones and type their answers directly into the computer. And interactive voice response (IVR) technology can be used during telephone interviews, whereby respondents hear prerecorded audio renderings of the questions and type their answers on their telephone keypads. Thus, computer assisted self-administered interviewing (CASAI) and Audio CASAI afford all of the advantages of computerized face-to-face and telephone interviewing, along with many of the advantages of self-administration (for a review of modes, see Groves et al., 2009).

Smartphones and other mobile devices are now also being used to collect survey data. Although small screens pose challenges for the presentation of

questions and the recording of answers, mobile devices offer the advantage of allowing researchers to know some respondents' physical locations at the time the questionnaire is completed and allow respondents to take and send real-time photographs to researchers.

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 of representative samples of the general population are much more expensive than telephone interviews, which are about as expensive as Internet and paper-and-pencil surveys of representative samples of general populations these days.

THE POPULATION. Several 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, writing or computer operation. Thus, this mode of data collection is inappropriate if a non-negligible portion of the population being studied does not meet this minimum literacy proficiency. Motivation is another relevant factor – when researchers suspect that respondents may be unmotivated to participate in a survey, or to read questions carefully, interviewers are typically more effective at eliciting participation than are paper or email invitations. 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; Groves, Cialdini, & Couper, 1992; Marquis, Cannell, & Laurent, 1972).

SAMPLING STRATEGY. The sampling strategy to be used may sometimes suggest a particular mode of data collection. For example, some pre-election polling organizations draw their samples from lists of currently registered voters. Such lists often provide only names and mailing addresses and no phone numbers for many people; this limits the mode of data collection to face-to-face interviewing or mailed questionnaire self-administration.

DESIRED RESPONSE RATE. Face-to-face surveys routinely achieve the highest response rates, especially when conducted by the federal government. Telephone surveys typically achieve lower response rates, and Internet surveys typically achieve even lower response rates. Self-administered mail surveys can achieve high response rates if they follow an extensive protocol (Dillman, 1978; Dillman, Smyth, & Christian, 2008).

QUESTION FORM. If a survey includes open-ended questions that require probing to clarify the details of respondents' answers, face-to-face or telephone interviewing is 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; Kreuter, Presser, & Tourangeau, 2008; Newman, Des Jarlais, Turner, Gribble, Cooley, & Paone, 2002; Wiseman, 1972).

QUESTIONNAIRE LENGTH. Face-to-face data collection is thought to permit the longest continuous interviews – an hour or more – without respondent break-offs midway through. Telephone interviews are typically quite a bit shorter, usually lasting no more than 30 minutes, because respondents are often uncomfortable staying on the phone for longer.

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 and Internet surveys can be completed in very little time, within a matter of days or even hours.

AVAILABILITY OF STAFF AND FACILITIES. Self-administered mail surveys require the fewest facilities

and can be completed by a small staff. Face-to-face and telephone surveys typically require much larger staffs, including interviewers, their supervisors, and coordinators of the supervisors. Telephone surveys can be conducted from a central location with sufficient office space to accommodate a staff of interviewers, but such interviewing can also be done from interviewers' homes via a central computer system that allows supervisors in a central location to monitor the interviewers' computers and conversations.

DATA QUALITY. A number of studies have compared the accuracy of data collected in various different modes. To date, they suggest that face-to-face interviewing may yield more representative samples and more accurate and honest reports than do telephone interviews (e.g., Holbrook, Green, & Krosnick, 2003). Computer self-administration also appears to elicit more accurate data than does telephone interviewing (e.g., Chang & Krosnick, 2009, 2010; Yeager et al., 2011). Nonetheless, telephone interviewing remains remarkably accurate in assessments of accuracy, such as predicting the outcomes of national elections.

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 (J. M. Converse & Schuman, 1974; Fowler & Mangione, 1986, 1990; Lavrakas, 2010; 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. Third, interviewers are responsible for executing the survey in a standardized way. The second and third responsibilities may sometimes 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 in which interviewer race may have had an impact along these lines involved the 1989 Virginia gubernatorial race. Pre-election 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 attributable at least in part to social desirability. Some survey respondents 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, although no race-of-interviewer effects appeared on numerous items that did not explicitly ask about liking of blacks (see also Anderson, Silver, & Abramson, 1988; Cotter, Cohan, & Coulter, 1983; Davis, 1997; Davis & Silver, 2003; Hyman, Feldman, & Stember, 1954; Schaeffer, 1980). 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 (Billiet & Loosveldt, 1988; Fowler & Mangione, 1986, 1990; Hansen, 2007; Lavrakas, 1993). Careful interviewer

training can presumably reduce random and systematic survey error resulting from 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 two 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

1. how to use all interviewing equipment;
2. procedures for randomly selecting respondents within households;
3. techniques for eliciting survey participation and avoiding refusals;
4. opportunities to gain familiarity with the survey instrument and to practice administering the questionnaire;
5. instructions regarding how and when to probe incomplete responses;
6. instructions on how to record answers to open- and closed-ended questions; and
7. 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). It is important that the training session involve supervised practice interviewing. Pairs of trainees are routinely asked to take turns playing the roles of interviewer and respondent. Such role playing might also involve the use of various "respondent scripts" that present potential problems for the interviewer to practice handling. Interviewers can be trained both in how to recruit potential respondents and in how to ask the survey's questions.

SUPERVISION. Carefully monitoring ongoing data collection permits early detection of problems and seems likely to improve data quality. 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, to identify potential problems. Researchers can also monitor the data collected by interviewers in real time, to be sure that open-ended answers are being transcribed properly, for example (cf., Steve et al., 2008).

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 make audio recordings of some or all 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 some interviewers to falsify some of the questionnaires that they turn in, and some occasionally do. This is referred to as curbstoning, a topic addressed in a 2009 report by the American Association for Public Opinion Research, the nation's leading professional association of survey researchers (see http://www.aapor.org/Content/aapor/AdvocacyandInitiatives/StandardsandEthics/InterviewerFalsificationPracticesandPolicies/ReporttoAAPORStandardsCommonInterviewerFalsification/default.htm). 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). This can only be accomplished if contact information for respondents is maintained by the researcher, which must be done carefully in order to keep identities confidential and never connected to responses.

TOTAL SURVEY ERROR

As is no doubt obvious by now, high-quality survey data collection can be very costly. And many survey researchers, even those with big budgets, nonetheless have limits on how much they can spend on a project. Such financial limitations routinely force researchers to make choices about how to spend their money. That is, a decision to spend money on one component of a study (e.g., paying financial incentives) will necessarily mean that less money is available to spend on other aspects of the data collection effort. For example, in order to be able to pay for an extra day of interviewer

training, a researcher might have to reduce the number of interviewer hours available for conducting the survey's interviews.

How should a researcher go about making these choices? What principles should guide the allocation of resources? Building on the work of Hansen (e.g., Hansen & Madow, 1953), the "total survey error" perspective suggests that such decisions should be made in ways that maximize the accuracy of the obtained data (cf. Dillman, 1978, Fowler, 1988; Groves, 1989). The total survey error perspective is based on the notion 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 seven major components: coverage error, sampling error, nonresponse error, specification error, measurement error, adjustment error, and processing error. *Coverage error* refers to the bias that can result when the pool of potential respondents 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 members of a sample. *Specification error* refers to how well the constructs the researchers purport to have assessed were actually measures. 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 respondents to respond in one way or another), and the questionnaire (e.g., ambiguous or confusing wording, biased question wording or response options). *Adjustment error* refers to the statistical corrections that have been made to address issues such as unequal probabilities of selection at the time of sampling, and the problems that noncoverage and nonresponse may have caused in yielding a final sample that is not representative of the target population. *Processing error* refers to how well the "raw dataset" has been cleaned and processed (e.g., the coding of open-ended verbatim transcripts, the transformation of data gathered at the interval level into categorical variables, the creation of multi-item scales, etc.) in creating a final data set that researchers will analyze.

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 seven. There is remarkably little scientific evidence available at the moment to quantify the gains in data accuracy that result from a dollar being spent in various different ways, so researchers are currently left to make guesses about the benefits of various types of expenditures. Perhaps in the future, more empirical research will be done to guide these sorts of decisions.

CONCLUSIONS

This chapter offers only the very beginning of an introduction to the process of survey data collection, and interested readers can turn to more extensive treatments of these issues to gain further mastery (e.g., Groves, Fowler, Couper, Lepkowski, Singer, & Tourangeau, 2009; Lavrakas, 2008). We hope here to have illustrated for social psychologists just how complex and challenging the survey data collection process is and why it is worth the trouble. Social science, after all, is meant to gain insights into populations of people, and the organizations that fund our work deserve to know whether our findings accurately describe entire populations or describe only narrow subsets. Given the strong traditions of representative sampling in other disciplines, such as political science, sociology, and economics, psychologists place themselves at a disadvantage if they completely ignore the scientific imperative to confirm the generalizability and applicability of their findings in rigorous ways. Surveys offer the opportunity to do just this.

But even for a social-personality psychologist who chooses to forego studies of representative samples, the survey methodology literature has a lot to offer to help that person do his or her work more effectively. Specifically, the huge and growing literature on questionnaire design offers guidelines for optimizing measurement in laboratory experiments of convenience samples. To ignore that literature is to risk using measuring tools that acquire a great deal of random or nonrandom measurement error and therefore make it more difficult for a researcher to detect real relationships between variables. Therefore, in the interest of minimizing the number of respondents in a study while maximizing the ability to gauge effect sizes accurately, questions should be designed

according to the principles evolving in the survey questionnaire design arena. Just one striking example of suboptimality is social-personality psychologists' continued reliance on agree-disagree rating scales in the face of the huge literature documenting the damaging impact of acquiescence response bias. We look forward to improved measurement by social psychologists and enhanced efficiency of the scientific inquiry that is likely to result, and such changes can be spurred by careful attention to the findings of survey methodologists on questionnaire design.

Furthermore, it is beyond dispute that the diversity of life experiences, perspectives, and approaches to decision making is much, much greater in the entire adult population than in the subpopulation of students enrolled in college psychology courses. Therefore, surveys of general population samples offer not only the opportunity to produce findings legitimating generalization, but also exciting opportunities for theory development. By collecting and analyzing data from large, heterogeneous full-population samples, researchers can be spurred to consider a wide range of new moderator variables that may encapsulate the impact of life situations and individual attributes on social psychological processes. And as a result, our theories may end up being richer, especially thanks to the emergence of new technologies that permit collection of many sorts of data in the course of daily life from representative samples of people.

But even if occasional use of survey data from representative samples does not change the nature of our scientific findings at all, it seems likely that use of such data will (appropriately) enhance the perceived credibility of our enterprise and will illustrate that social psychologists are willing to invest the effort and funds necessary to move beyond convenient laboratory studies of captive audiences in order to objectively evaluate the applicability of our claims. This, in and of itself, seems like sufficient justification for social psychologists to learn about how survey data are collected, learn about how to analyze them properly, make funding requests that are sufficient to allow such work, and enrich our work product as a result.

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.

Alderman, H., Behrman, J. R., Kohler, H., Maluccio, J. A., & Watkins, S. C. (2001). Attrition in longitudinal household survey data. *Demographic Research, 5*(4), 79–124.

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 R. M. Hauser, D. Mechanic, A. O. Haller, & T. S. Hauser (Eds.), *Sociological structure and behavior: Essays in honor of William Hamilton Sewell* (pp. 311–329). New York: Academic Press.

Alwin, D. F., & Krosnick, (1991). The reliability of survey attitude measurement. The influence of question and respondent attributes. *Sociological Methods and Research, 20*, 139–181.

Anderson, B., Silver, B., & Abramson, P. (1988). The effects of the race of the interviewer on measures of electoral participation by blacks in SRC national election studies. *Public Opinion Quarterly, 52*, 53–83.

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.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*, 287–311.

Becketti, S., Gould, W., Lillard, L., & Welch, F. (1988). The PSID after fourteen years: An evaluation. *Journal of Labor Economics, 6*(4), 472–492.

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.

Berinsky, A. J. (1999). The two faces of public opinion. *American Journal of Political Science, 43*, 1209–1230.

Billiet, J., & Loosveldt, G. (1988). Improvement of the quality of responses to factual survey questions by interviewer training. *Public Opinion Quarterly, 52*, 190–211.

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.

Blumberg, S. J., & Luke, J. V. (2012). *Wireless substitution: Early release of estimates from the National Health Interview Survey, January–June 2012*. Atlanta, GA: Centers for Disease Control and Prevention.

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.

Box-Steffensmeier, J. M., Jacobson, G. C., & Grant, J. T. (2000). Question wording and the house vote choice: Some experimental evidence. *Public Opinion Quarterly*, *64*, 257–270.

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

Cantor, D., O'Hare, B., & O'Connor, K. (2008). The use of monetary incentives to reduce non-response in random digit dial telephone surveys. In J. M. Lepkowski et al. (Eds.), *Advances in telephone survey methodology* (pp. 471–498). New York: Wiley.

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.

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing vs. the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*, 641–678.

Chang, L., & Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized questionnaires: An experiment. *Public Opinion Quarterly*, *74*, 154–167.

Chanley, V. A., Rudolph, T. J., & Rahn, W. M. (2000). The origins and consequences of public trust in government. A time series analysis. *Public Opinion Quarterly*, *64*, 239–256.

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

Clinton, J. D. (2001). *Panel bias from attrition and conditioning: A case study of the knowledge networks panel*. Stanford, CA: Stanford University Press.

Clinton, J. D., & Rogers, S. (2012). *Robo-polls: Taking cues from traditional sources*? Unpublished manuscript, Vanderbilt University, Nashville, TN.

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.

Cook, C., Heath, F., & Thompson, R. L. (2001). Score reliability in web- or Internet-based surveys: Unnumbered graphic rating scales versus Likert-type scales. *Educational and Psychological Measurement*, *61*, 697–706.

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.

Cotter, P., Cohen, J. & Coulter, P. (1982). Race-of-interviewer effects in telephone interviews. *Public Opinion Quarterly*, *46*, 278–284.

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

Couper, M. (2008). *Designing effective web surveys*. Cambridge: Cambridge University Press.

Couper, M. P., Tourangeau, R., Conrad, R. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review*, *24*, 227–245.

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly*, *65*, 230–253.

Courneya, K. S., Jones, L. W., Rhodes, R. E., & Blanchard, C. M. (2003). Effect of response scales on self-reported exercise frequency. *American Journal of Health Behavior*, *27*, 613–622.

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.

Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, *64*, 413–428.

Davies, T., & Gangadharan, S. P. (Eds.). (2009). *Online deliberation: Design, research, and practice*. Stanford, CA: CSLI Publications.

Davis, D. W. (1997). Nonrandom measurement error and race of interviewer effects among African Americans. *Public Opinion Quarterly*, *61*, 183–207.

Davis, D. W., & Silver, B. D. (2003). Stereotype threat and race of interviewer effects in a survey on political knowledge. *American Journal of Political Science*, *47*, 33–45.

DeBell, M., & Krosnick, J. A. (2009). *Computing weights for American National Election Study survey data*. ANES Technical Report series, no. nes012427. Ann Arbor, MI, and Palo Alto, CA: American National Election Studies. Retrieved August 26, 2013, from http://www.electionstudies.org/resources/papers/nes012427.pdf.

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., Smyth, J.D., & Christian, L. M. (2008). *Internet, mail, and mixed-mode surveys: The tailored design method*. New York: Wiley.

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

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

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 Personality and Social Psychology*, *37*, 621–634.

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

Eveland, Jr., W. P., Hayes, A. F., Shah, D. V., & Kwak, N. (2005). Understanding the relationship between communication and political knowledge: A model comparison approach using panel data. *Political Communication*, *22*, 423–446.

Falaris, E. M., & Peters, H. E. (1998). Survey attrition and schooling choices. *The Journal of Human Resources*, *33*, 531–554.

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.

Fitzgerald, J., Gottschalk, P., & Moffitt, R. (1998a). An analysis of sample attrition in panel data: The Michigan panel study of income dynamics. *NBER Technical Working Papers,* National Bureau of Economic Research, Inc.

Fitzgerald, J., Gottschalk, P., & Moffitt, R. (1998b). An analysis of the impact of sample attrition on the second generation of respondents in the Michigan panel study of income dynamics. *The Journal of Human Resources*, *33*, 300–344.

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. (2009). *Survey research methods* (4th ed.). Thousand Oaks, CA: Sage.

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

Fowler, Jr., F. J., & 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.

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, *72*, 892–913.

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.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029–1046.

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.

Green, D. P., & Gerber, A. S. (2006). Can registration-based sampling improve the accuracy of midterm election forecasts? *Public Opinion Quarterly*, *70*, 197–223.

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. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*, 646–675.

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, *56*, 475–495.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. New York: Wiley.

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

Haddock, G., & Zanna, M. P. (1998). On the use of open-ended measures to assess attitudinal components. *British Journal of Social Psychology*, *37*, 129–149.

Hamamura, T., Meijer, Z., Heine, S. J., Kamaya, K., & Hori, I. (2009). Approach-avoidance motivation and information processing: A cross-cultural analysis. *Personality and Social Psychology Bulletin*, *35*, 454–462.

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, K. M. (2007). The effects of incentives, interview length, and interviewer characteristics on response rates in CATI-study. *International Journal of Public Opinion Research*, *19*, 112–121.

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

Harzing, A.-W., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A. et al. (2009). Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International Business Review*, *18*, 417–432.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems, 44* (2), 174–199.

Heckathorn, D. D. (2002). Respondent-driven sampling II: Deriving valid estimates from chain-referral samples of hidden populations. *Social Problems*, *49*(1), 11–34.

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.

Hess, J., Singer, E., & Bushery, J. (1999). Predicting test-retest reliability from behavior coding. *International Journal of Public Opinion Research*, *11*, 346–360.

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.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone vs. face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67*, 79–125.

Holbrook A. L., Krosnick, J. A., Moore, D., & Tourangeau, R. (2007). Response order effects in dichotomous categorical questions presented orally: The impact of questiona and respondent attributes. *Public Opinion Quarterly*, *71*, 325–348.

Holbrook, A. L., Krosnick, J. A., & Pfent, A. M. (2008). The causes and consequences of response rates in surveys by the news media and government contractor survey research firms. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. De Leeuw, L. Japec, P. J. Lavrakas, M. W. Link, & R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 499–528). New York: Wiley.

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.

Hurd, M. D. (1999). Anchoring and acquiescence bias in measuring assets in household surveys. *Journal of Risk and Uncertainty*, *19*, 111−136.

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

Jackman, M. R. (1973). 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.

Kam, C. D., & Ramos, J. M. (2008). Joining and leaving the rally. *Public Opinion Quarterly*, *72*, 619–650.

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

Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S., 2000; Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, *64*,125–148.

Kellstedt, P. M., Peterson, D. A. M., & Ramirez, M. D. (2010). The macro politics of gender gap. *Public Opinion Quarterly*, *74*, 477–498.

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.

Kim, S.-H., Scheufele, D. A., Shanahan, J., & Choi, D.-H. (2011). Deliberation in spite of controversy? News media and the public's evaluation of a controversial issue in South Korea. *Journalism & Mass Communication Quarterly*, *88*, 2320–2336.

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.

Kitayama, S., Park, H., Sevincer, A. T., Karasawa, M., & Uskul, A. K. (2009). A cultural task analysis of implicit independence: Comparing North America, Western Europe, and East Asia. *Journal of Personality and Social Psychology*, *97*, 236–255.

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.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*, 847–865.

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. (forthcoming). *Designing great questionnaires: Insights from psychology*. New York: Oxford University Press.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Ruud, P. A., Smith, V. K., Moody, W. R., Green, M. C., & Conaway, M. (2002). The impact of "no opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, *66*, 371–403.

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.

Lavrakas, P. J. (Ed.). (2008). *Encyclopedia of survey research methods*. Thousand Oaks, CA: Sage.

Lavrakas, P. J. (2010). Telephone surveys. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research*. San Diego, CA: Elsevier.

Li, X. (2008). Third-person effect, optimistic bias, and sufficiency resource in Internet use. *Journal of Communication*, *58*, 568–587.

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.

Link, M. W., Battaglia, M. P., Frankel, M. R., Osborn, L., & Mokdad, A. H. (2007). Reaching the U.S. cell phone generation: Comparison of cell phone survey results with an ongoing landline telephone survey. *Public Opinion Quarterly*, *71*, 814–839.

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

Lozano, L. M., Garcia-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, *4*(2), 73−79.

Luskin, R. C., & Bullock, J. G. (2011). "Don't know" means "don't know": DK responses and the public's level of political knowledge. *Journal of Politics*, *73*, 547–557.

Mann, C. B. (2005). Unintentional voter mobilization: Does participation in preelection surveys increase voter turnout? *The Annals of the American Academy of Political and Social Science*, *601*, 155–168.

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.

McFarland, S. G. (1981). Effects of question order on survey responses. *Public Opinion Quarterly*, *45*, 208–215.

Merkle, D., & Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.). *Survey Nonresponse* (pp. 243–258). New York: Wiley.

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

Miller, J. M., & Krosnick, J. A. (1998). The impact of candidate name order on election outcomes. *Public Opinion Quarterly*, *62*, 291–330.

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.

Moore, D. W. (2002). New types of question-order effects: Additive and subtractive. *Public Opinion Quarterly*, *66*, 80–91.

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.), *Lifespan 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).

Newman, J. C., Des Jarlais, D. C., Turner, C. F., Gribble, J., Cooley, P., & Paone, D. (2002). The differential effects of face-to-face and computer interview modes. *American Journal of Public Health*, *92*, 294–297.

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.

Office of Information and Regulatory Affairs. (2006). Questions and answers when designing surveys for information collections. Washington, DC: Office of Management and Budget. Retrieved August 26, 2013, from http://www.whitehouse.gov/sites/default/files/omb/inforeg/pmc_survey_guidance_2006.pdf.

Pasek, J. (2012a). Package "anesrake." Retrieved August 26, 2013, from http://cran.r-project.org/web/packages/anesrake/anesrake.pdf.

Pasek, J. (2012b). Online weighting tool. Retrieved August 26, 2013, from http://joshpasek.com/category/software/

Pasek, J., Tahk, A., Lelkes, Y., Krosnick, J. A., Payne, K., Akhtar, O., & Tompson, T. (2009). Determinants of turnout and candidate choice in the 2008 U.S. Presidential election: Illuminating the impact of racial prejudice and other considerations. *Public Opinion Quarterly*, *73*, 943–994.

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, B. K., Krosnick, J. A., Pasek, J., Lelkes, Y., Akhtar, O., & Tompson, T. (2010). Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology*, *46*, 367–374.

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

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

Petty, R. E., & Cacioppo, J. T. (1996). Addressing disturbing and disturbed consumer behavior: Is it necessary to change the way we conduct behavioral science? *Journal of Marketing Research*, *33*, 1–8.

Presser, S. (1990). Measurement issues in the study of social change. *Social Forces*, *68*, 856–868.

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.

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., & Singer, E. (2004). *Methods for testing and evaluating survey questionnaires*. New York: Wiley-Interscience.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1–15.

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.

Richardson, J. D. (2004). Isolating frequency scale effects on self-reported loneliness. *Personality and Individual Differences*, *36*, 235–244.

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

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: construct validation of a single-item measure and the Rosenberg self-esteem scale. *Personality and Social Psychology Bulletin*, *27*, 151–161.

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.

Salganik, M. J., & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, *34*(1), 193–239.

Sapsford, R. (2007). *Survey research* (2nd ed.). London: Sage.

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

Saris, W. E. (1998). Ten years of interviewing without interviewer: The Telepanel. In M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nichols, & J. M. O'Reilly (Eds.), *Computer assisted survey information collection*. New York: John Wiley and Sons.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, NJ: John Wiley & Sons.

Saris, W., Revilla, M., Krosnick, J. A., & Shaeffer, E. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, *4*, 61–79.

Schaeffer, N. C. (1980). Evaluating race-of-interviewer effects in a national survey. *Sociological Methods and Research*, *8*, 400–419.

Schuman, H. (2008). *Method and meaning in polls and surveys*. Cambridge, MA: Harvard University Press.

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., & Presser, S. (1996). *Questions and answers in attitude surveys.* New York: Academic Press.

Schuman, H., & Scott, J. (1989). Response effects over time: Two experiments. *Sociological Methods and Research*, *17*, 398–408.

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.

Schwarz, N., Hippler, H., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, *49*, 388–395.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, F. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 618–630.

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.

Shen, F., Wang, N., Guo, Z., & Guo, L. (2009). Online network size, efficacy, and opinion expression: Assessing the impacts of Internet use in China. *International Journal of Public Opinion Research*, *21*, 451–476.

Singer, E., Van Hoewyk, J., & Maher, M. P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, *64*, 171–188.

Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The Annals of the American Academy of Political and Social Science*, *645*, 112–141.

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.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Mcbride, M. (2009). Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, *73*, 325–337.

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.

Steve, K. W., Burks, A. T., Lavrakas, P. J., Brown, K. D., & Hoover, J. B. (2008). Monitoring telephone interviewer performance. In J. M. Lepkowski, C. Turner, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas, M. W. Link, & R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 201–422). New York: Wiley.

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, Jr., O. T., & 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.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

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.

Trzesniewski, K. M., Donnellan, B., & Lucas, R. E. (Eds.). (2010). *Secondary data analysis: An introduction for psychologists*. Washington, DC: American Psychological Association.

Trzesniewski, K. M., Donnellan, M. B., & Robins, R. W. (2003). Stability of self-esteem across the life span. *Journal of Personality and Social Psychology*, *84*, 205–220.

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

United Nations Economic and Social Commission for Asia and the Pacific. (1999a). Computer Assisted Personal Interviewing (CAPI). In *Guidelines on the application of new technology to population data collection and capture*. New York: The United Nations. Retrieved August 25, 2013, from http://www.unescap.org/stat/pop-it/pop-guide/capture_ch03.pdf.

United Nations Economic and Social Commission for Asia and the Pacific. (1999b). Computer Assisted Telephone Interviewing (CATI). In *Guidelines on the application of new technology to population data collection and capture*. New York: The United Nations. Retrieved August 26, 2013, from http://www.unescap.org/stat/pop-it/pop-guide/capture_ch04.pdf.

Van De Walle, S., & Van Ryzin, G. G. (2011). The order of questions in a survey on citizen satisfaction with public services: Lessons from a split-ballot experiment. *Public Administration*, *89*, 1436–1450.

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.

Voogt, R. J. J., & Van Kempen, H. (2002). Nonresponse bias and stimulus effects in the Dutch National Election Study. *Quality and Quantity*, *36*, 325–345.

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

Watson, D. (2003). Sample attrition between waves 1 and 5 in the European Community Household Panel. *European Sociological Review*, *19*, 361–378.

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.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*, 236–247.

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.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*, 956–972.

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.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications.

Wilson, D. C. (2010). Perceptions about the amount of interracial prejudice depend on racial group membership *and* question order. *Public Opinion Quarterly*, *74*, 344–356.

Winkler, J. D., Kanouse, D. E., & Ware, Jr., J. E. (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.

Wojcieszak, M. E. (2012). On strong attitudes and group deliberation: Relationships, structure, changes, and effects. *Political Psychology*, *33*, 225–242.

Wright, S. D., Middleton, R. T., & Yon, R. (2012). The effect of racial group consciousness on the political participation of African-Americans and Black ethnics in Miami-Dade County, Florida. *Political Research Quarterly*, *65*, 629–641.

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

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and Internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, *75*, 709–747.

Zabel, J. E. (1998). An analysis of attrition in the panel study of income dynamics and the survey of income and program participation with an application to a model of labor market behavior. *The Journal of Human Resources*, *33*, 479–506.

Zagorsky, J., & Rhoton, P. (1999). *Attrition and the national longitudinal survey's women cohorts*. Manuscript, Center for Human Resource Research, Ohio State University.

Ziliak, J. P., & Kniesner, T. J. (1998). The importance of sample attrition in life cycle labor supply estimation. *Journal of Human Resources*, *33*, 507–530.