

CHAPTER 3

OPTIMIZING
SURVEY
QUESTIONNAIRE
DESIGN IN
POLITICAL
SCIENCE:
INSIGHTS FROM
PSYCHOLOGY

JOSH PASEK

JON A. KROSNICK

QUESTIONNAIRES have long been a primary means of gathering data on political behavior (F. H. Allport 1940; G. W. Allport 1929; Campbell et al. 1960; Dahl 1961; Lazarsfeld and Rosenberg 1949–1950; Merriam 1926; Woodward and Roper 1950). Many of the most frequently studied and important measurements made to understand mass political action have been done with questions in the American National Election Studies (ANES) surveys and other such data collection

enterprises. Although in principle, it might seem desirable to observe political behavior directly rather than relying on people's descriptions of it, questionnaire-based measurement offers tremendous efficiencies and conveniences for researchers over direct observational efforts. Furthermore, many of the most important explanatory variables thought to drive political behavior are subjective phenomena that can only be measured via people's descriptions of their own thoughts. Internal political efficacy, political party identification, attitudes toward social groups, trust in government, preferences among government policy options on specific issues, presidential approval, and many more such variables reside in citizens' heads, so we must seek their help by asking them to describe those constructs for us.

A quick glance at ANES questionnaires might lead an observer to think that the design of self-report questions need follow no rules governing item format, because formats have differed tremendously from item to item. Thus, it might appear that just about any question format is as effective as any other format for producing valid and reliable measurements. But in fact, this is not true. Nearly a century's worth of survey design research suggests that some question formats are optimal, whereas others are suboptimal.

In this chapter, we offer a summary of this literature's suggestions. In doing so, we point researchers toward question formats that appear to yield the highest measurement reliability and validity. Using the American National Election Studies as a starting point, the chapter illuminates general principles of good questionnaire design, desirable choices to make when designing new questions, biases in some question formats and ways to avoid them, and strategies for reporting survey results. Finally, the chapter offers a discussion of strategies for measuring voter turnout in particular, as a case study that poses special challenges. We hope that the tools we present will help scholars to design effective questionnaires and utilize self-reports so that the data gathered are useful and the conclusions drawn are justified.

THE QUESTIONS WE HAVE ASKED

Many hundreds of questions have been asked of respondents in the ANES surveys, usually more than an hour's worth in one sitting, either before or after a national election. Many of these items asked respondents to place themselves on rating scales, but the length of these scales varies considerably. For example, some have 101 points, such as the feeling thermometers:

Feeling Thermometer. I'd like to get your feelings toward some of our political leaders and other people who are in the news these days. I'll read the name of a person and I'd

like you to rate that person using something we call the feeling thermometer. The feeling thermometer can rate people from 0 to 100 degrees. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the person. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the person. Rating the person at the midpoint, the 50 degree mark, means you don't feel particularly warm or cold toward the person. If we come to a person whose name you don't recognize, you don't need to rate that person. Just tell me and we'll move on to the next one. (ANES 2004)

Other ratings scales have offered just seven points, such the ideology question:

Liberal-conservative Ideology. We hear a lot of talk these days about liberals and conservatives. When it comes to politics, do you usually think of yourself as extremely liberal, liberal, slightly liberal; moderate or middle of the road, slightly conservative, conservative, extremely conservative, or haven't you thought much about this? (ANES 2004)

Still others have just five points:

Attention to Local News about the Campaign. How much attention do you pay to news on local news shows about the campaign for President—a great deal, quite a bit, some, very little, or none? (ANES 2004)

Or three points:

Interest in the Campaigns. Some people don't pay much attention to political campaigns. How about you? Would you say that you have been very much interested, somewhat interested or not much interested in the political campaigns so far this year? (ANES 2004)

Or just two:

Internal efficacy. Please tell me how much you agree or disagree with these statements about the government: "Sometimes politics and government seem so complicated that a person like me can't really understand what's going on." (ANES 2004)

Whereas the internal efficacy measure above offers generic response choices ("agree" and "disagree"), which could be used to measure a wide array of constructs, other items offer construct-specific response alternatives (meaning that the construct being measured is explicitly mentioned in each answer choice), such as:

Issue Importance. How important is this issue to you personally? Not at all important, not too important, somewhat important, very important, or extremely important? (ANES 2004)

Some rating scales have had verbal labels and no numbers on all the points, as in the above measure of issue importance, whereas other rating scales have numbered points with verbal labels on just a few, as in this case:

Defense Spending. Some people believe that we should spend much less money for defense. Suppose these people are at one end of the scale, at point number 1. Others feel that defense spending should be greatly increased. Suppose these people are at the other end, at point 7.

And, of course, some other people have opinions somewhere in between at points 2, 3, 4, 5 or 6. Where would you place yourself on this scale, or haven't you thought much about this? (ANES 2004)

In contrast to all of the above closed-ended questions, some other questions are asked in open-ended formats:

Candidate Likes–dislikes. Is there anything in particular about Vice President Al Gore that might make you want to vote for him?

Most Important Problems. What do you think are the most important problems facing this country?

Political Knowledge. Now we have a set of questions concerning various public figures. We want to see how much information about them gets out to the public from television, newspapers and the like. What job or political office does Dick Cheney now hold?

(ANES 2004)

Some questions offered respondents opportunities to say they did not have an opinion on an issue, as in the ideology question above (“or haven't you thought much about this?”). But many questions measuring similar constructs do not offer that option, such as:

U.S. Strength in the World. Turning to some other types of issues facing the country. During the past year, would you say that the United States' position in the world has grown weaker, stayed about the same, or has it grown stronger?

Variations in question design are not, in themselves, problematic. Indeed, one cannot expect to gather meaningful data on a variety of issues simply by altering a single word in a “perfect,” generic question. To that end, some design decisions in the ANES represent the conscious choices of researchers based on pre-testing and the literature on best practices in questionnaire design. In many cases, however, differences between question wordings are due instead to the intuitions and expectations of researchers, a desire to retain consistent questions for time-series analyses, or researchers preferring the ease of using an existent question rather than designing and pre-testing a novel one.

All of these motivations are understandable, but there may be a better way to go about questionnaire design to yield better questions. Poorly designed questions can produce momentary confusion among respondents or more widespread frustration and small compromises in reliability or large and systematic biases in measurement or analysis results. Designing optimal measurement tools in surveys sometimes requires expenditure or more resources (by asking longer questions or more questions to measure a single construct), but many measurements can be made optimal simply by changing wording without increasing a researcher's costs. But to do so requires understanding the principles of optimal design, which we review next.

BASIC DESIGN PRINCIPLES

Good questionnaires are easy to administer, yield reliable data, and accurately measure the constructs for which the survey was designed. When rapid administration and acquiring reliable data conflict, however, we lean toward placing priority on acquiring accurate data. An important way to enhance measurement accuracy is to ask questions that respondents can easily interpret and answer and that are interpreted similarly by different respondents. It is also important to ask questions in ways that motivate respondents to provide accurate answers instead of answering sloppily or intentionally inaccurately. How can we maximize respondent motivation to provide accurate self-reports while minimizing the difficulty of doing so? Two general principles underlie most of the challenges that researchers face in this regard. They involve (1) understanding the distinction between “optimizing” and “satisficing,” and (2) accounting for the conversational framework that shapes the survey response process. We describe these theoretical perspectives next.

Optimizing and Satisficing

Imagine the ideal survey respondent, whom we'll call an optimizer. Such an individual goes through four stages in answering each survey question (though not necessarily strictly sequentially). First, the optimizer reads or listens to the question and attempts to discern the question's intent (e.g., “the researcher wants to know how often I watch television programs about politics”). Second, the optimizer searches his or her memory for information useful to answer the question (e.g., “I guess I usually watch television news on Monday and Wednesday nights for about an hour at a time, and there's almost always some political news covered”). Third, the optimizer evaluates the available information and integrates that information into a summary judgment (e.g., “I watch two hours of television about politics per week”). Finally, the optimizer answers the question by translating the summary judgment onto the response alternatives (e.g. by choosing “between 1 and 4 hours per week”) (Cannell et al. 1981; Krosnick 1991; Schwarz and Strack 1985; Tourangeau and Rasinski 1988; Turner and Martin 1984).

Given the substantial effort required to execute all the steps of optimizing when answering every question in a long questionnaire, it is easy to imagine that not every respondent implements all of the steps fully for every question (Krosnick 1999; Krosnick and Fabrigar 1998). Indeed, more and more research indicates that some individuals sometimes answer questions using only the most readily available information, or, worse, look for cues in the question that point toward easy-to-select answers and choose them so as to do as little thinking as possible (Krosnick 1991). The act of abridging the search for information or skipping

it altogether is termed “survey satisficing” and appears to pose a major challenge to researchers (Krosnick 1991; Simon 1957). When respondents satisfice, they give researchers answers that are at best loosely related to the construct of interest and may sometimes be completely unrelated to it.

Research on survey satisficing has revealed a consistent pattern of who satisfices and when. Respondents are likely to satisfice when the task of answering a particular question optimally is difficult, when the respondent lacks the skills needed to answer optimally, or when he or she is unmotivated (Krosnick 1991; Krosnick and Alwin 1987). Hence, satisficers are individuals who have limited cognitive skills, fail to see sufficient value in a survey, find a question confusing, or have simply been worn down by a barrage of preceding questions (Krosnick 1991; Krosnick, Narayan, and Smith 1996; McClendon 1986, 1991; Narayan and Krosnick 1996). These individuals tend to be less educated and score lower on psychological batteries like “need for cognition” than non-satisficers (Narayan and Krosnick 1996). Importantly, they do not represent a random subset of the population, and they tend to satisfice in systematic, rather than stochastic, ways. Hence, to ignore satisficers is to introduce potentially problematic bias in survey results.

No research has yet identified a surefire way to prevent respondents from satisficing, but a number of techniques for designing questions and putting them together into questionnaires seem to reduce the extent to which respondents satisfice (Krosnick 1999). Questions, therefore, should be designed to minimize the incentives to satisfice and maximize the efficiency of the survey for optimizers.

Conversational Norms and Conventions

In most interpersonal interactions, participants expect a conversant to follow certain conversational standards. When people violate these conversational norms and rules, confusion and misunderstandings often ensue. A “cut” is a very different thing when requested at a butcher shop or at a hair stylist, for instance. Without context, it is difficult to know what is being sought by a question such as, “What sorts of cuts do you offer?” From this perspective, a variety of researchers have attempted to identify the expectations that conversants bring to conversations, so any potentially misleading expectations can be overcome. In his seminal work *Logic and Conversation*, Grice (1975) proposed a set of rules that speakers usually follow and listeners usually assume that speakers follow: that they should be truthful, meaningfully informative, relevant, and to the point. This perspective highlights a critical point that survey researchers often ignore: respondents enter all conversations with expectations, and when researchers violate those expectations (which they often do unwittingly), measurement accuracy can be compromised (Lipari 2000; Schuman and Ludwig 1983; Schwarz 1996).

Krosnick, Li, and Lehman (1990) illustrated the impact of conversational norms. They found the form in which information was presented in a survey question

could substantially change how respondents answered. In everyday conversations, when people list a series of pieces of information leading to a conclusion, they tend to present that which they think of as the most important information last. When Krosnick et al.'s respondents were given information and were asked to make decisions with that information, the respondents placed more weight on the information that was presented last. In another study, Holbrook et al. (2000) presented response options to survey questions in either a normal ("are you *for* or *against* X?") or unusual ("are you *against* or *for* X?") order. Respondents whose question used the normal ordering were quicker to respond to the questions and answered more validly. Thus, breaking rules of conversation manipulates and compromises the quality of answers.

Implications

Taken together, these theoretical perspectives suggest that survey designers should try to follow three basic rules. Surveys should:

- (1) be designed to make questions as easy as possible for respondents to answer,
- (2) take steps to discourage satisficing, and
- (3) be sure not to violate conversational conventions without explicitly saying so, to avoid confusion and misunderstandings.

The specifics of how to accomplish these three goals are not always obvious. Cognitive pre-testing (which involves having respondents restate questions in their own words and think aloud while answering questions, to highlight misunderstandings that need to be prevented) is always a good idea, but many of the specific decisions that researchers must make when designing questions can be guided by the findings of past studies on survey methodology. The literature in these areas, reviewed below, provides useful and frequently counter-intuitive answers.

DESIGNING OPTIMAL SURVEY QUESTIONS

Open-ended questions or closed questions?

In the 1930s and 1940s, when modern survey research was born, a debate emerged as to whether researchers should ask open-ended questions or should ask respondents to select among a set of offered response choices (J. M. Converse 1984). Each method had apparent benefits. Open-ended questions could capture the exact sentiments of individuals on an issue with any tones of nuance and without the

possibility for answer choices to color respondent selection. Closed questions seemed easier to administer and to analyze, and more of them could be asked in a similar amount of time (Lazarsfeld 1944). Perhaps more out of convenience than merit, closed questions eclipsed open-ended ones in contemporary survey research. For example, in surveys done by major news media outlets, open-ended questions constituted a high of 33 percent of questions in 1936 and dropped to 8 percent of questions by 1972 (T. Smith 1987).

The administrative ease of closed questions, however, comes with a distinct cost. Respondents tend to select among offered answer choices rather than selecting an “other, specify” (Belson and Duncan 1962; Bishop et al. 1988; Lindzey and Guest 1951; Oppenheim 1966; Presser 1990b). If every potential option is offered by a question, then this concern is irrelevant. For most questions, however, offering every possible answer choice is not practical. And when some options are omitted, respondents who would have selected them choose among the offered options, thereby changing the distribution of responses as compared to what would have been obtained if a complete list had been offered (Belson and Duncan 1962). Therefore, an open-ended format would be preferable in this sort of situation.

Open-ended questions also discourage satisficing. When respondents are given a closed question, they might settle for choosing an appropriate-sounding answer. But open-ended questions demand individuals to generate an answer on their own and do not point respondents toward any particular response, thus demanding more thought and consideration (Oppenheim 1966). Furthermore, many closed questions require respondents to answer an open-ended question in their minds first (e.g., “what is the most important problem facing the country?”) and then to select the answer choice that best matches that answer. Skipping the latter, matching step will make the respondent’s task easier and thereby encourage optimizing when answering this and subsequent questions.

Closed questions can also present particular problems when seeking numbers. Schwarz et al. (1985) manipulated response alternatives for a question gauging amount of television watching and found striking effects. When “more than 2½ hours” was the highest category offered, only 16 percent of individuals reporting watching that much television. But when five response categories broke up “more than 2½ hours” into five sub-ranges, nearly 40 percent of respondents placed themselves in one of those categories. This appears to occur because whatever range is in the middle of the set of offered ranges is perceived to be typical or normal by respondents, and this implicit message sent by the response alternatives alters people’s reports (Schwarz 1995). Open-ended questions seeking numbers do not suffer from this potential problem.

The higher validity of open-ended questions does not mean that every question should be open-ended. Open-ended questions take longer to answer and must be systematically coded by researchers (Lazarsfeld 1944). When the full spectrum of possible responses is known, closed questions are an especially appealing

alternative. But when the full spectrum of answers is not known, or when a numeric quantity is sought (e.g., “during the last month, how many times did you talk to someone about the election?”), open-ended questions are preferable.

- Before asking a closed question seeking categorical answers, however, researchers should pre-test an open-ended version of the question on the population of interest, to be sure the offered list of response alternatives is comprehensive.

Rating Questions or Ranking Questions?

Rating questions are very common in surveys (e.g., the “feeling thermometer” and “internal efficacy” questions above). Such questions are useful because they place respondents on the continua of interest to researchers and are readily susceptible to statistical analysis. Furthermore, rating multiple items of a given type can permit comparisons of evaluations across items (McIntyre and Ryans 1977; Moore 1975; Munson and McIntyre 1979).

In some situations, researchers are interested in obtaining rank order of objects from respondents (e.g., rank these candidates from most desirable to least desirable). In such situations, asking respondents to rank-order the objects is an obvious measurement option, but it is quite time-consuming (Munson and McIntyre 1979). Therefore, it is tempting to ask respondents instead to rate the objects individually and to derive a rank order from the ratings.

Unfortunately, though, rating questions sometimes entail a major challenge: when asked to rate a set of objects on the same scale, respondents sometimes fail to differentiate their ratings, thus clouding analytic results (McCarty and Shrum 2000). This appears to occur because some respondents choose to satisfice by non-differentiating: drawing a straight line down the battery of rating questions. For example, in one study with thirteen rating scales, 42 percent of individuals evaluated nine or more of the objects identically (Krosnick and Alwin 1988). And such non-differentiation is most likely to occur under the conditions that foster satisficing (see Krosnick 1999).

If every object deserves an identical rating, rating scales would not be problematic. But when researchers are interested in understanding how respondents rank-order objects when forced to do so, satisficing-induced non-differentiation in ratings quickly devolves into misleading data (Alwin and Krosnick 1985). Fortunately, respondents can be asked to rank candidates or objects instead. Although ranking questions take more time, rankings acquire responses that are less distorted by satisficing and are more reliable and valid than ratings (Alwin and Krosnick 1985; Krosnick and Alwin 1988; Miethe 1985; Reynolds and Jolly 1980).

- Thus, ranking questions are preferable for assessing rank orders of objects.

Rating Scale Points

Although the ANES “feeling thermometer” measure has been used in numerous American National Election Study surveys, it has clear and obvious drawbacks: the meanings of the many scale points are not clear and uniformly interpreted by respondents. Only nine of the points are labeled with words on the show-card handed to respondents, and a huge proportion of respondents choose one of those nine points (Weisberg and Miller 1979). But labeling is not the only problem with long rating scales like the feeling thermometer. When points are unlabeled, individuals gravitate toward multiples of five or ten, similarly ignoring most of the scale (Schaeffer and Bradburn 1989). Furthermore, subjective differences in interpreting response alternatives may mean that one person’s 80 is equivalent to another’s 65 (Wilcox, Siegelman, and Cook 1989). Therefore, this very long and ambiguous rating scale introduces considerable error into analysis.

Although 101 points is far too many for a meaningful scale, providing only two or three response choices for a rating scale can make it impossible for respondents to provide evaluations at a sufficiently refined level to communicate their self-perceptions (Alwin 1992; Alwin and Krosnick 1985). Too few response alternatives can provide a particular challenge for optimizers who are attempting to map complex opinions onto limited answer choices. A large body of research has gone into assessing the most effective range of options to offer respondents (Alwin 1992; Alwin and Krosnick 1985; Cox 1980; Lissitz and Green 1975; Lodge and Tursky 1979; Matell and Jacoby 1972; Ramsay 1973; Schuman and Presser 1981).

- Ratings tend to be more reliable and valid when five points are offered for unipolar dimensions (e.g., “not at all important” to “extremely important”; Lissitz and Green 1975) and seven points for bipolar dimensions (e.g., “Dislike a great deal” to “like a great deal” Green and Rao 1970).

Another drawback of the “feeling thermometer” scale is its numerical scale point labels. Labels are meant to improve respondent interpretation of scale points, but the meanings of most of the numerically labeled scale points are unclear. It is therefore preferable to put verbal labels on all rating scale points to clarify their intended meanings, which increases the reliability and validity of ratings (Krosnick and Berent 1993). Providing numeric labels in addition to the verbal labels increases respondents’ cognitive burden but does not increase data quality and in fact can mislead respondents about the intended meanings of the scale points (e.g., Schwarz et al. 1991). Verbal labels with meanings that are not equally spaced from one another can cause respondent confusion (Klockars and Yamagishi 1988), so the selected verbal labels should have equally spaced meanings (Hofmans et al. 2007; Schwarz, Grayson, and Knauper 1998; Wallsten et al. 1986).

“Don’t Know” Options and Attitude Strength

Although some questionnaire designers advise that opinion questions offer respondents the opportunity to say they do not have an opinion at all (e.g., Vaillancourt 1973), others do not advise including “don’t know” or “no opinion” response options (Schuman and Presser 1981). And most major survey research firms have routinely trained their interviewers to probe respondents when they say “don’t know” to encourage them to offer a substantive answer instead. The former advice is sometimes justified by claims that respondents may sometimes be unfamiliar with the issue in question or may not have enough information about it to form a legitimate opinion (e.g., P. Converse 1964). Other supportive evidence has shown that people sometimes offer opinions about extremely obscure or fictitious issues, thus showing that they are manufacturing non-attitudes instead of confessing ignorance (e.g., Bishop, Tuchfarber, and Oldendick 1986; Hawkins and Coney 1981; Schwarz 1996).

In contrast, advice to avoid offering “don’t know” options is justified by the notion that such options can encourage satisficing (Krosnick 1991). Consistent with this argument, when answering political knowledge quiz questions, respondents who are encouraged to guess after initially saying “don’t know” tend to give the correct answer at better-than-chance rates (Mondak and Davis 2001). Similarly, candidate preferences predict actual votes better when researchers discourage “don’t know” responses (Krosnick et al. 2002; Visser et al. 2000). Thus, discouraging “don’t know” responses collects more valid data than does encouraging such responses. And respondents who truly are completely unfamiliar with the topic of a question will say so when probed, and that answer can be accepted at that time, thus avoiding collecting measurements of non-existent “opinions.”

- Thus, because many people who initially say “don’t know” do indeed have a substantive opinion, researchers are best served by discouraging these responses in surveys.

Converse (1964) did have an important insight, though. Not all people who express an opinion hold that view equally strongly, based upon equal amounts of information and thought. Instead, attitudes vary in their strength. A strong attitude is very difficult to change and has powerful impact on a person’s thinking and action. A weak attitude is easy to change and has little impact on anything. To understand the role that attitudes play in governing a person’s political behavior, it is valuable to understand the strength of those attitudes. Offering a “don’t know” option is not a good way to identify weak attitudes. Instead, it is best to ask follow-up questions intended to diagnose the strength of an opinion (see Krosnick and Abelson 1992).

Acquiescence Response Bias

In everyday conversations, people want to be agreeable and want to be agreed with (Brown and Levinson 1987). In surveys, however, when researchers ask questions, they mean to invite all possible responses, even when asking respondents whether they agree or disagree with a statement offered by a question. “Likert scales” is the label often used to describe the agree–disagree scales that are used in many surveys these days. Such scales are appreciated by both designers and respondents because they speed up the interview process. Unfortunately, though, respondents are biased toward agreement. Some 10–20 percent of respondents tend to agree with both a statement and its opposite when the direction of agreement is reversed (e.g., Schuman and Presser 1981). This tendency toward agreeing is known as acquiescence response bias, and may occur for a variety of reasons. First, conversational conventions dictate that people should be agreeable and polite (Bass 1956; Campbell et al. 1960). Second, people tend to defer to individuals of higher authority (a position they assume the researcher holds) (Carr 1971; Lenski and Leggett 1960). Additionally, when inclined to satisfice, agreeing with a statement is more likely than disagreeing (Krosnick 1991).

Whatever the cause, acquiescence presents a major challenge for researchers (Bass 1955). Consider, for example, the ANES question measuring internal efficacy. If certain respondents are more likely to agree with any statement regardless of its content, then these individuals will appear to believe that government and politics are too complicated to understand, even if that is not their view. And any correlations between this questions and other questions could be due to associations with the individual’s actual internal efficacy or his or her tendency to acquiesce (Kenski and Jomini 2004; Wright 1975).

Agree–disagree rating scales are extremely popular in social science research, and researchers rarely take steps to minimize the impact of acquiescence on research findings (Zuckerman et al. 1995). One such step is to balance batteries of questions, such that affirmative answers indicate a high level of the construct for half the items and a low level of the construct for the other half, thus placing acquiescers at the midpoint of the final score’s continuum (Bass 1956; Cloud and Vaughan 1970). Unfortunately, this approach simply moves acquiescers from the agree of a rating scale (where they don’t necessarily belong) to the midpoint of the final score’s continuum (where they also don’t necessarily belong) (Billiet and McClendon 2000).

A more effective solution becomes apparent when we recognize first that answering an agree–disagree question is more cognitively demanding than answering a question that offers construct-specific response alternatives. This is so because in order to answer most agree–disagree questions (e.g., “Sometimes politics is so complicated that I can’t understand it”), the respondent must answer a construct-specific version of it in his or her own mind (“How often is politics so complicated

that I can't understand it?") and then translate the answer onto the agree–disagree response continuum. And in this translation process, a person might produce an answer that maps onto the underlying construct in a way the researcher would not anticipate. For example, a person might disagree with the statement, "Sometimes politics is so complicated that I can't understand it," either because politics is never that complicated or because politics is always that complicated. Thus, the agree–disagree continuum would not be monotonically related to the construct of interest. For all these reasons,

- it is preferable simply to ask questions with construct-specific response alternatives.

Yes/No questions and True/False questions are also subject to acquiescence response bias (Fritzley and Lee 2003; Schuman and Presser 1981). In these cases, a simple fix involves changing the question so that it explicitly offers all possible views. For example, instead of asking "Do you think abortion should be legal?" one can ask "Do you think abortion should or should not be legal?"

Response Order Effects

Another form of satisficing is choosing the first plausible response option one considers, which produces what are called response order effects (Krosnick 1991, 1999; Krosnick and Alwin 1987). Two types of response order effects are primacy effects and recency effects. Primacy effects occur when respondents are inclined to select response options presented near the beginning of a list (Belson 1966). Recency effects occur when respondents are inclined to select options presented at the end of a list (Kalton, Collins, and Brook 1978). When categorical (non-rating scale) response options are presented visually, primacy effects predominate. When categorical response options are presented orally, recency effects predominate. When rating scales are presented, primacy effects predominate in both the visual and oral modes. Response order effects are most likely to occur under the conditions that foster satisficing (Holbrook et al. 2007).

One type of question that can minimize response order effects is the seemingly open-ended question (SOEQ). SOEQs separate the question from the response alternatives with a short pause to encourage individuals to optimize. Instead of asking, "If the election were held today, would you vote for Candidate A or Candidate B?" response order effects can be reduced by asking, "If the election were held today, whom would you vote for? Would you vote for Candidate A or Candidate B?" The pause after the question and before the answer choices encourages respondents to contemplate, as if when answering an open-ended question, and then offers the list of the possible answers to respondents (Holbrook et al. 2007).

- By rotating response order or using SOEQs, researchers can prevent the order of the response options from coloring results.

Response order effects do not only happen in surveys. They occur in elections as well. In a series of natural experiments, Brook and Upton (1974), Krosnick and Miller (1998), Koppell and Steen (2004), and others found consistent patterns indicating that a few voters choose the first name on the ballot, giving that candidate an advantage of about 3 percent on average. Some elections are decided by less than 3 percent of the vote, so name order can alter an election outcome. When telephone survey questions mirror the name order on the ballot, those surveys are likely to manifest a recency effect, which would run in the direction opposite to what would be expected in the voting booth, thus creating error in predicting the election outcome. Many survey firms rotate candidate name order to control for potential effects, but this will yield accurate forecasts only in states, such as Ohio, that rotate candidate name order in voting booths.

Question Order Effects

In 1948, one survey asked Americans whether Communist reporters should be allowed in the United States and found that the majority (63 percent) said “no.” Yet in another survey, an identical question found 73 percent of Americans believed that Communist reporters should be allowed. This discrepancy turned out to be attributable to the impact of the question that preceded the target question in the latter survey. In the later experiment, a majority of Americans said “yes” when the item immediately followed a question about whether American reporters should be allowed in Russia. Wanting to appear consistent and attuned to the norm of even-handedness after hearing the initial question, respondents were more willing to allow Communist reporters into the US (Schuman and Presser 1981).

A variety of other types of question order effects have been identified. *Subtraction* occurs when two nested concepts are presented next to each other (e.g., George W. Bush and the Republican Party) as items for evaluation. When a question about the Republican Party follows a question about George W. Bush, respondents assume that the questioner does not want them to include their opinion of Bush in their evaluations of the GOP (Schuman, Presser, and Ludwig 1981). *Perceptual contrast* occurs when one rating follows another, and the second rating is made in contrast to the first one. For example, respondents who dislike George Bush may be inclined to offer a more favorable rating of John McCain when a question about McCain follows Bush than when the question about McCain is asked first (Schwarz and Bless 1992; Schwarz and Strack 1991). And *priming* occurs when questions earlier in the survey increase the salience of certain attitudes or beliefs in the mind

of the respondent (e.g., preceding questions about abortion may make respondents more likely to evaluate George W. Bush based on his abortion views) (Kalton et al. 1978). Also, asking questions later in a long survey enhances the likelihood that respondents will satisfice (Krosnick 1999).

Unfortunately, it is rarely easy to discern question order effects. Rotating the order of questions across respondents might seem sensible, but doing so may yield topics of questions that seem to jump around in ways that don't seem obviously sensible and tax respondents' memories (Silver and Krosnick 1991). And rotating question order will not make question order effects disappear. Therefore, the best researchers can do is to use past research on question order effects as a basis for being attentive to possible question order effects in a new questionnaire.

Attitude Recall

It would be very helpful to researchers if respondents could remember the opinions they held at various times in the past and describe them accurately in surveys. Unfortunately, this is rarely true. People usually have no recollection of how they thought about things at previous times. When asked, they will happily guess, and their guesses are strongly biased—people tend to assume they always believed what they believe today (Goethals and Reckman 1973; Roberts 1985). Consequently, attitude recall questions tend to produce wildly inaccurate results (T. Smith 1984). Because of the enormous amount of error and bias associated with these questions, they cannot be used for statistical analyses. Instead, attitude change must be assessed prospectively.

- Only by measuring attitudes at multiple time points is it possible to gain an accurate understanding of attitude change.

The Danger of Asking “Why?”

Social science spends much of its time determining causality. Instead of running dozens of studies and spending millions of dollars, it might seem much more efficient simply to ask people to describe the reasons for their thoughts and actions (Lazarsfeld 1935). Unfortunately, respondents rarely know why they think and act as they do (Nisbett and Wilson 1977; E. R. Smith and Miller 1978; Wilson and Dunn 2004; Wilson and Nisbett 1978). People are happy to guess when asked, but their guesses are rarely informed by any genuine self-insight and are usually no more accurate than would be guesses about why someone else thought or acted as they did.

- Consequently, it is best not to ask people to explain why they think or act in particular ways.

Social Desirability

Some observers of questionnaire data are skeptical of their value because they suspect that respondents may sometimes intentionally lie in order to appear more socially admirable, thus manifesting what is called social desirability response bias. Many observers have attributed discrepancies between survey rates of voter turnout and official government turnout figures to intentional lying by survey respondents (Belli, Traugott, and Beckmann 2001; Silver, Anderson, and Abramson 1986). Rather than appearing not to fulfill their civic duty, some respondents who did not vote in an election are thought to claim that they did so. Similar claims have been made about reports of illegal drug use and racial stereotyping (Evans, Hansen, and Mittelmark 1977; Sigall and Page 1971).

A range of techniques have been developed to assess the scope of social desirability effects and to reduce the likelihood that people's answers are distorted by social norms. These methods either assure respondents that their answers will be kept confidential or seek to convince respondents that the researcher can detect lies—making it pointless not to tell the truth (Paulhus 1984). Interestingly, although these techniques have often revealed evidence of social desirability response bias, the amount of distortion is generally small. Even for voting, where social desirability initially seemed likely to occur, researchers have found slightly lower voting rates in some surveys but no large universal effect (Abelson, Loftus, and Greenwald 1992; Duff et al. 2007; Holbrook and Krosnick in press; Presser 1990a). Even after controlling for social desirability, surveyed turnout rates remained well above those reported in government records.

A number of other errors are likely to contribute to the overestimate in voter turnout. First, official turnout records contain errors, and those errors are more likely to be omissions of individuals who did vote than inclusions of individuals who did not vote. A significant number of voting records get misplaced (Presser, Traugott, and Traugott 1990). Second, many individuals who are "potential" voters fall outside of survey sampling frames (Clausen 1968–1969; McDonald 2003; McDonald and Popkin 2001). Third, individuals who choose not to participate in a political survey are less likely to vote than individuals who do participate (Burden 2000; Clausen 1968–1969). Fourth, individuals surveyed before an election may be made more likely to vote as the result of the interview experience (Kraut and McConahay 1973; Traugott and Katosh 1979). Surveys like the ANES could overestimate turnout partially because follow-up surveys are conducted with individuals who had already been interviewed (Clausen 1968–1969). All of these factors contribute to the unrealistic expectation that survey results should match published voter turnout figures.

Another reason for apparent overestimation of turnout may be acquiescence, because answering "yes" to a question about voting usually indicates having done so (Abelson, Loftus, and Greenwald 1992). Second, respondents who regularly vote

may not recall that, in a specific instance, they failed to do so (Belli, Traugott, and Beckmann 2001; Belli, Traugott, and Rosenstone 1994; Belli et al. 1999). Each of these alternate proposals has gotten some empirical support. So although social desirability may be operating, especially in telephone interviews, it probably accounts for only a small portion of the over-reporting effect.

Question Wording

Although much of questionnaire design should be considered a science rather than an art (J. M. Converse and Presser 1986; Krosnick and Fabrigar 1998), the process of selecting words for a question is thought to be artistic and intuitive (Payne 1951). A question's effectiveness can easily be undermined by long, awkward wording that taps multiple constructs. Despite the obvious value of pithy, easy-to-understand queries, questionnaire designers regularly provide tome-worthy introductions. One obvious example is the preamble for the "feeling thermometer." When tempted to use such a long and complicated introduction, researchers should recognize the problematic nature of the question. Indeed, the "feeling thermometer" is handicapped in multiple ways, as we have outlined above. More generally, it is unwise to ask leading questions laden with jargon or ambiguity.

Choices of words for questions are worth agonizing over, because even very small changes can produce sizable differences in responses. In one study, for instance, 73 percent of respondents said they strongly or somewhat "favored" policies on average, whereas only 45 percent strongly or somewhat "supported" the same policies (Krosnick 1989). Many studies have produced similar findings, showing that differences in word choice can change individuals' responses remarkably (e.g., Rugg 1941). But this does not mean that respondents are arbitrary or fickle. The choice of a particular word or phrase can change the perceived meaning of a question in sensible ways and therefore change the judgment that is reported.

- Therefore, researchers should be very careful to select words tapping the exact construct they mean to measure.

Conclusion

Numerous studies of question construction suggest a roadmap of best practices. Systematic biases caused by satisficing and the violation of conversational conventions can distort responses, and researchers have both the opportunity and ability to minimize those errors. These problems therefore are mostly those of design. That is, they can generally be blamed on the researcher, not on the respondent. And fortunately, intentional lying by respondents appears to be very rare and

preventable by using creative techniques to assure anonymity. So again, accuracy is attainable.

The American National Election Study is a smorgasbord of some good and many suboptimal questions. Despite these shortcomings, those survey questions nonetheless offer a window into political attitudes and behaviors that would be impossible to achieve through any other research design. Nonetheless, scholars designing their own surveys should not presume that previously written ANES questions are the best ones to use. Applying best practices in questionnaire design will yield more accurate data and more accurate substantive findings about the nature and origins of mass political behavior.

REFERENCES

- ABELSON, R. P., LOFTUS, E., and GREENWALD, A. G. 1992. Attempts to Improve the Accuracy of Self-Reports of Voting. In *Questions About Questions: Inquiries into the Cognitive Bases of Surveys*, ed. J. M. Tanur. New York: Russell Sage Foundation.
- ALLPORT, F. H. 1940. Polls and the Science of Public Opinion. *Public Opinion Quarterly*, 4/2: 249–57.
- ALLPORT, G. W. 1929. The Composition of Political Attitudes. *American Journal of Sociology*, 35/2: 220–38.
- ALWIN, D. F. 1992. Information Transmission in the Survey Interview: Number of Response Categories and the Reliability of Attitude Measurement. *Sociological Methodology*, 22: 83–118.
- and KROSNICK, J. A. 1985. The Measurement of Values in Surveys: A Comparison of Ratings and Rankings. *Public Opinion Quarterly*, 49/4: 535–52.
- AMERICAN NATIONAL ELECTION STUDIES 2004. The 2004 National Election Study [codebook]. Ann Arbor, MI: University of Michigan, Center for Political Studies (producer and distributor). <<http://www.electionstudies.org>>.
- BASS, B. M. 1955. Authoritarianism or Acquiescence? *Journal of Abnormal and Social Psychology*, 51: 616–23.
- . 1956. Development and evaluation of a scale for measuring social acquiescence. *Journal of Abnormal and Social Psychology*, 53/3: 296–9.
- BELLI, R. F., TRAUOGOTT, M. W., and BECKMANN, M. N. 2001. What Leads to Voting Overreports? Contrasts of Overreporters to Validated Voters And Admitted Nonvoters in the American National Election Studies. *Journal of Official Statistics*, 17/4: 479–98.
- BELLI, R. F., TRAUOGOTT, S., and ROSENSTONE, S. J. 1994. Reducing Over-Reporting of Voter Turnout: An Experiment Using a “Source Monitoring” Framework. In *NES Technical Reports*.
- TRAUOGOTT, M. W., YOUNG, M., and MCGONAGLE, K. A. 1999. Reducing Vote Over-reporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring. *Public Opinion Quarterly*, 63/1: 90–108.
- BELSON, W. A. 1966. The Effects of Reversing the Presentation Order of Verbal Rating Scales. *Journal of Advertising Research*, 6: 30–7.

- and DUNCAN, J. A. 1962. A Comparison of the Check-List and The Open Response Questioning Systems. *Applied Statistics*, 11/2: 120–32.
- BILLIET, J. B., and MCCLENDON, M. J. 2000. Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items. *Structural Equation Modeling*, 7/4: 608–28.
- BISHOP, G. F., HIPPLER, H.-J., SCHWARZ, N., and STRACK, F. 1988. A Comparison of Response Effects in Self-Administered and Telephone Surveys. In *Telephone Survey Methodology*, ed. R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nocholls II, and J. Waksberg. New York: Wiley.
- TUCHFARBER, A. J., and OLDENDICK, R. W. 1986. Opinions on Fictitious Issues: The Pressure to Answer Survey Questions. *Public Opinion Quarterly*, 50/2: 240–50.
- BROOK, D., and UPTON, G. J. G. 1974. Biases in Local Government Elections Due to Position on the Ballot Paper. *Applied Statistics*, 23/3: 414–19.
- BROWN, P., and LEVINSON, S. C. 1987. *Politeness: Some Universals in Language Usage*. New York: Cambridge University Press.
- BURDEN, B. C. 2000. Voter Turnout and the National Election Studies. *Political Analysis*, 8/4: 389–98.
- CAMPBELL, A., CONVERSE, P. E., MILLER, W. E., and STOKES, D. 1960. *The American Voter: Unabridged Edition*. New York: Wiley.
- CANNELL, C. F., MILLER, P. V., and OSKENBERG, L. 1981. Research on Interviewing Techniques. *Sociological Methodology*, 12: 389–437.
- CARR, L. G. 1971. The Srole Items and Acquiescence. *American Sociological Review*, 36/2: 287–93.
- CLAUSEN, A. R. 1968–1969. Response Validity: Vote Report. *Public Opinion Quarterly*, 32/4: 588–606.
- CLOUD, J., and VAUGHAN, G. M. 1970. Using Balanced Scales to Control Acquiescence. *Sociometry*, 33/2: 193–202.
- CONVERSE, J. M. 1984. Strong Arguments and Weak Evidence: the Open/Closed Questioning Controversy of the 1940s. *Public Opinion Quarterly*, 48/1: 267–82.
- and PRESSER, S. 1986. *Survey Questions: Handcrafting the Standard Questionnaire*. Ed. M. S. Lewis-Beck (138 vols.), Volume 63. Thousand Oaks, Calif.: Sage.
- CONVERSE, P. E. 1964. The Nature of Belief Systems in Mass Publics. In *Ideology and Discontent*, ed. D. Apter. New York: Free Press.
- COX, E. P., III. 1980. The Optimal Number of Response Alternatives for a Scale: A Review. *Journal of Marketing Research*, 17/4: 407–22.
- DAHL, R. A. 1961. The Behavioral Approach in Political Science: Epitaph for a Monument to a Successful Protest. *American Political Science Review*, 55/4: 763–72.
- DUFF, B., HANMER, M. J., PARK, W.-H., and WHITE, I. K. 2007. Good Excuses: Understanding Who Votes With An Improved Turnout Question. *Public Opinion Quarterly*, 71/1: 67–90.
- EVANS, R. I., HANSEN, W. B., and MITTELMARK, M. B. 1977. Increasing the Validity of Self-Reports of Smoking Behavior in Children. *Journal of Applied Psychology*, 62/4: 521–3.
- FRITZLEY, V. H., and LEE, K. 2003. Do Young Children Always Say Yes to Yes–No Questions? A Metadevelopmental Study of the Affirmation Bias. *Child Development*, 74/5: 1297–313.
- GOETHALS, G. R., and RECKMAN, R. F. 1973. Perception of Consistency in Attitudes. *Journal of Experimental Social Psychology*, 9: 491–501.
- GREEN, P. E., and RAO, V. R. 1970. Rating Scales and Information Recovery. How Many Scales and Response Categories to Use? *Journal of Marketing*, 34/3: 33–9.

- GRICE, H. P. 1975. Logic and Conversation. In *Syntax and Semantics*, Volume 3, *Speech Acts*, ed. P. Cole and J. L. Morgan. New York: Academic Press.
- HAWKINS, D. I., and CONEY, K. A. 1981. Uninformed Response Error in Survey Research. *Journal of Marketing Research*, 18/3: 370–4.
- HOFMANS, J., THEUNS, P., BAEKELANDT, S., MAIRESSE, O., SCHILLEWAERT, N., and COOLS, W. 2007. Bias and Changes in Perceived Intensity of Verbal Qualifiers Effected by Scale Orientation. *Survey Research Methods*, 1/2: 97–108.
- HOLBROOK, A. L., and KROSINICK, J. A. In Press. Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique. *Public Opinion Quarterly*.
- CARSON, R. T, and MITCHELL, R. C. 2000. Violating Conversational Conventions Disrupts Cognitive Processing of Attitude Questions. *Journal of Experimental Social Psychology*, 36: 465–94.
- MOORE, D., and TOURANGEAU, R. 2007. Response Order Effects in Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes. *Public Opinion Quarterly*, 71/3: 325–48.
- KALTON, G., COLLINS, M., and BROOK, L. 1978. Experiments in Wording Opinion Questions. *Applied Statistics*, 27/2: 149–61.
- KENSKI, K, and JOMINI, N. 2004. The Reciprocal Effects of External and Internal Political Efficacy?: Results from the 2000 U.S. Presidential Election. In *World Association for Public Opinion Research*. Phoenix, Ariz.
- KLOCKARS, A. J., and YAMAGISHI, M. 1988. The Influence of Labels and Positions in Rating Scales. *Journal of Educational Measurement*, 25/2: 85–96.
- KOPPELL, J. G. S., and STEEN, J. A. 2004. The Effects of Ballot Position on Election Outcomes. *Journal of Politics*, 66/1: 267–81.
- KRAUT, R. E., and MCCONAHAY, J. B. 1973. How Being Interviewed Affects Voting: An Experiment. *Public Opinion Quarterly*, 37/3: 398–406.
- KROSINICK, J. A. 1989. Question Wording and Reports of Survey Results: The Case of Louis Harris and Aetna Life and Casualty. *Public Opinion Quarterly*, 53: 107–13.
- 1991. Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, 5: 213–36.
- 1999. Survey Research. *Annual Review of Psychology*, 50: 537–67.
- and ABELSON, R. P. (1992). The Case for Measuring Attitude Strength in Surveys. In *Questions About Questions: Inquiries into the Cognitive Bases of Surveys*, ed. J. M. Tanur. New York: Russell Sage Foundation.
- and ALWIN, D. F. 1987. An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51/2: 201–19.
- .1988. A Test of the Form-Resistant Correlation Hypothesis: Ratings, Rankings, and the Measurement of Values. *Public Opinion Quarterly*, 52/4: 526–38.
- and BERENT, M. K. 1993. Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format. *American Journal of Political Science*, 37/3: 941–64.
- and FABRIGAR, L. R. 1998. *Designing Good Questionnaires: Insights from Psychology*. New York: Oxford University Press.
- LI, F., and LEHMAN, D. R. 1990. Conversational Conventions, Order of Information Acquisition, and the Effect of Base Rates and Individuating Information on Social Judgments. *Journal of Personality and Social Psychology*, 59: 1140–52.

- and MILLER, J. M. 1998. “The Impact of Candidate Name Order on Election Outcomes.” *Public Opinion Quarterly* 62/3: 291–330.
- NARAYAN, S., and SMITH, W. R. 1996. Satisficing in Surveys: Initial Evidence. *New Directions for Program Evaluation*, 70: 29–44.
- 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., and 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.
- LAZARSFELD, P. F. 1935. The Art of Asking *Why*: Three Principles Underlying the Formulation of Questionnaires. *National Marketing Review*, 1: 32–43.
- 1944. The Controversy Over Detailed Interviews – An Offer for Negotiation. *Public Opinion Quarterly*, 8/1: 38–60.
- and ROSENBERG, M. 1949–1950. The Contribution of the Regional Poll to Political Understanding. *Public Opinion Quarterly*, 13/4: 569–86.
- LENSKI, G. E., and LEGGETT, J. C. 1960. Caste, Class, and Deference in the Research Interview. *American Journal of Sociology*, 65/5: 463–7.
- LINDZEY, G. E., and GUEST, L. 1951. To Repeat – Check Lists Can Be Dangerous. *Public Opinion Quarterly*, 15/2: 355–8.
- LIPARI, L. 2000. Toward a Discourse Approach to Polling. *Discourse Studies*, 2/2: 187–215.
- LISSITZ, R. W., and GREEN, S. B. 1975. Effect of the Number of Scale Points on Reliability: A Monte Carlo Approach. *Journal of Applied Psychology*, 60/1: 10–3.
- LODGE, M., and TURSKY, B. 1979. Comparisons between Category and Magnitude Scaling of Political Opinion Employing SRC/CPS Items. *American Political Science Review*, 73/1: 50–66.
- MATELL, M. S., and JACOBY, J. 1972. Is There an Optimal Number of Alternatives for Likert-scale Items? Effects of Testing Time and Scale Properties. *Journal of Applied Psychology*, 56/6: 506–9.
- MCCARTY, J. A., and SHRUM, L. J. 2000. The Measurement of Personal Values in Survey Research. *Public Opinion Quarterly*, 64/3: 271–98.
- MCCLENDON, M. J. 1986. Response-Order Effects for Dichotomous Questions. *Social Science Quarterly*, 67: 205–11.
- 1991. Acquiescence and Recency Response-Order Effects in Interview Surveys. *Sociological Methods & Research*, 20/1: 60–103.
- MCDONALD, M. P. 2003. On the Overreport Bias of the National Election Study Turnout Rate. *Political Analysis*, 11: 180–6.
- and POPKIN, S. L. 2001. The Myth of the Vanishing Voter. *American Political Science Review*, 95/4: 963–74.
- MCINTYRE, S. H., and RYANS, A. B. 1977. Time and Accuracy Measures for Alternative Multidimensional Scaling Data Collection Methods: Some Additional Results. *Journal of Marketing Research*, 14/4: 607–10.
- MERRIAM, C. E. 1926. Progress in Political Research. *American Political Science Review*, 20/1: 1–13.
- MIETHE, T. D. 1985. Validity and Reliability of Value Measurements. *Journal of Psychology*, 119/5: 441–53.
- MONDAK, J. J., and DAVIS, B. C. 2001. Asked and Answered: Knowledge Levels When We Will Not Take “Don’t Know” for an Answer. *Political Behavior*, 23/3: 199–222.

- MOORE, M. 1975. Rating Versus Ranking in the Rokeach Value Survey: An Israeli Comparison. *European Journal of Social Psychology*, 5/3: 405–8.
- MUNSON, J. M., and MCINTYRE, S. H. 1979. Developing Practical Procedures for the Measurement of Personal Values in Cross-Cultural Marketing. *Journal of Marketing Research*, 16/1: 48–52.
- NARAYAN, S., and KROSINICK, J. A. 1996. Education Moderates Some Response Effects in Attitude Measurement. *Public Opinion Quarterly*, 60/1: 58–88.
- NISBETT, R. E., and WILSON, T. D. 1977. Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review*, 84/3: 231–59.
- OPPENHEIM, A. N. 1966. *Questionnaire Design and Attitude Measurement*. New York: Basic Books.
- PAULHUS, D. L. 1984. Two-Component Models of Socially Desirable Responding. *Journal of Personality and Social Psychology*, 46/3: 598–609.
- PAYNE, S. L. 1951. *The Art of Asking Questions*. Princeton, N.J.: Princeton University Press.
- PRESSER, S. 1990a. Can Changes in Context Reduce Vote Overreporting in Surveys? *Public Opinion Quarterly*, 54/4: 586–93.
- 1990b. Measurement Issues in the Study of Social Change. *Social Forces*, 68/3: 856–68.
- TRAUGOTT, M. W., and TRAUGOTT, S. 1990. Vote “Over” Reporting in Surveys: The Records of the Respondents? In *International Conference on Measurement Errors*. Tucson, Ariz.
- RAMSAY, J. O. 1973. The Effect of Number of Categories in Rating Scales on Precision of Estimation of Scale Values. *Psychometrika*, 38/4: 513–32.
- REYNOLDS, T. J., and JOLLY, J. P. 1980. Measuring Personal Values: An Evaluation of Alternative Methods. *Journal of Marketing Research*, 17/4: 531–6.
- ROBERTS, J. V. 1985. The Attitude-Memory Relationship After 40 Years: A Meta-analysis of the Literature. *Basic and Applied Social Psychology*, 6/3: 221–41.
- RUGG, D. 1941. Experiments in Wording Questions: II. *Public Opinion Quarterly*, 5/1: 91–2.
- SCHAEFFER, N. C., and BRADBURN, N. M. 1989. Respondent Behavior in Magnitude Estimation. *Journal of the American Statistical Association*, 84 (406): 402–13.
- SCHUMAN, H., and LUDWIG, J. 1983. The Norm of Even-Handedness in Surveys as in Life. *American Sociological Review*, 48/1: 112–20.
- and PRESSER, S. 1981. *Questions and Answers in Attitude Surveys*. New York: Academic Press.
- and LUDWIG, J. 1981. Context Effects on Survey Responses to Questions About Abortion. *Public Opinion Quarterly*, 45/2: 216–23.
- SCHWARZ, N. 1995. What Respondents Learn from Questionnaires: The Survey Interview and the Logic of Conversation. *International Statistical Review*, 63/2: 153–68.
- 1996. *Cognition and Communication: Judgmental Biases, Research Methods and the Logic of Conversation*. Hillsdale, N.J.: Erlbaum.
- and BLESS, H. 1992. Scandals and the Public’s Trust in Politicians: Assimilation and Contrast Effects. *Personality and Social Psychology Bulletin*, 18/5: 574–9.
- GRAYSON, C. E., and KNAUPER, B. 1998. Formal Features of Rating Scales and the Interpretation of Question Meaning. *International Journal of Public Opinion Research*, 10/2: 177–83.
- and STRACK, F. 1985. Cognitive and Affective Processes in Judgments of Subjective Well-Being: A Preliminary Model. In *Economic Psychology*, ed. E. Kirchler and H. Brandstatter. Linz, Austria: R. Tauner.

- 1991. Context Effects in Attitude Surveys: Applying Cognitive Theory to Social Research. *European Review of Social Psychology*, 2: 31–50.
- HIPPLER, H.-J., DEUTSCH, B., and STRACK, F. 1985. Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments. *Public Opinion Quarterly*, 49/3: 388–95.
- KNAUPER, B., HIPPLER, H.-J., NOELLE-NEUMANN, E., and CLARK, L. 1991. Rating Scales: Numeric Values May Change the Meaning of Scale Labels. *Public Opinion Quarterly*, 55/4: 570–82.
- SIGALL, H., and PAGE, R. 1971. Current Stereotypes: A Little Fading, A Little Faking. *Journal of Personality and Social Psychology*, 18/2: 247–55.
- SILVER, B. D., ANDERSON, B. A., and ABRAMSON, P. R. 1986. Who Overreports Voting? *American Political Science Review*, 80/2: 613–24.
- SILVER, M. D., and KROSNICK, J. A. 1991. Optimizing Survey Measurement Accuracy by Matching Question Design to Respondent Memory Organization. In *Federal Committee on Statistical Methodology Conference*. NTIS: PB2002-100103. <<http://www.fcsm.gov/oipapers/Krosnick.pdf>>.
- SIMON, H. A. 1957. *Models of Man: Social and Rational*. New York: Wiley.
- SMITH, E. R., and MILLER, F. D. 1978. Limits on Perception of Cognitive Processes: A Reply to Nisbett and Wilson. *Psychological Review*, 85/4: 355–62.
- SMITH, T. W. 1984. Recalling Attitudes: An Analysis of Retrospective Questions on the 1982 GSS. *Public Opinion Quarterly*, 48/3: 639–49.
- 1987. The Art of Asking Questions, 1936–1985. *Public Opinion Quarterly*, 51/2: S95–108.
- TOURANGEAU, R., and RASINSKI, K. A. 1988. Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin*, 103/3: 299–314.
- TRAUGOTT, M. W., and KATOSH, J. P. 1979. Response Validity in Surveys of Voting Behavior. *Public Opinion Quarterly*, 43/3: 359–77.
- TURNER, C. F., and MARTIN, E. 1984. *Surveying Subjective Phenomena 1*. New York: Russell Sage Foundation.
- VAILLANCOURT, P. M. 1973. Stability of Children's Survey Responses. *Public Opinion Quarterly*, 37/3: 373–87.
- VISSE, P. S., KROSNICK, J. A., MARQUETTE, J. F., and CURTIN, M. F. 2000. Improving Election Forecasting: Allocation of Undecided Respondents, Identification of Likely Voters, and Response Order Effects. In *Election Polls, the News Media, and Democracy*, ed. P. Lavarakas and M. W. Traugott. New York: Chatham House.
- WALLSTEN, T. S., BUDESCU, D. V., RAPOPORT, A., ZWICK, R., and FORSYTH, B. 1986. Measuring the Vague Meanings of Probability Terms. *Journal of Experimental Psychology: General*, 115/4: 348–65.
- WEISBERG, H. F., and MILLER, A. H. 1979. *Evaluation of the Feeling Thermometer: A Report to the National Election Study Board Based on Data from the 1979 Pilot Survey*. ANES Pilot Study Report No. nes002241.
- WILCOX, C., SIEGELMAN, L., and COOK, E. 1989. Some Like It Hot: Individual Differences in Responses to Group Feeling Thermometers. *Public Opinion Quarterly*, 53/2: 246–57.
- WILSON, T. D., and DUNN, E. W. 2004. Self-Knowledge: It's Limits, Value, and Potential for Improvement. *Annual Review of Psychology*, 55: 493–518.
- and NISBETT, R. E. 1978. The Accuracy of Verbal Reports About the Effects of Stimuli on Evaluations and Behavior. *Social Psychology*, 41/2: 118–131.

- WOODWARD, J. L., and ROPER, E. 1950. Political Activity of American Citizens. *American Political Science Review*, 44/4: 872–85.
- WRIGHT, J. D. 1975. Does Acquiescence Bias the “Index of Political Efficacy?” *Public Opinion Quarterly*, 39/2: 219–26.
- ZUCKERMAN, M., KNEE, C. R., HODGINS, H. S., and MIYAKE, K. 1995. Hypothesis Confirmation: The Joint Effect of Positive Test Strategy and Acquiescence Response Set. *Journal of Personality and Social Psychology*, 68/1: 52–60.