(1999). Measures of Political Attitudes. (Ed.) John P. Robinson, Phillip R. Shaver, and Lawrence S. Wrightsman. San Diego, CA: Academic Press.

CHAPTER 2

Maximizing Questionnaire Quality

Jon A. Krosnick

Measurement tools are tremendously important in science because they are the lenses through which we see the world. To the biologist, reality is observed through a microscope. To the astronomer, learning takes place through a telescope. Geographers analyze photographs taken by satellites far from the earth's surface. Regardless of whether the measuring instrument is on a desk in front of the investigator or thousands of miles away, if there is a scratch or a spec of dust on the lens or if it is miscalibrated, a scientific analysis can end up far from the truth.

To many, and perhaps most, social scientists, measurement occurs via questionnaires. And most often our goal is to place people on measurement continua, whether they range from strong liberal to strong conservative, or from strong internationalist to strong isolationist, or from strongly pro–legalized abortion to strongly anti–legalized abortion. Just as for biologists, astronomers, and geographers, if our measuring instruments are flawed or miscalibrated, we can be seriously misled as well, placing people in the wrong places on a continuum relative to one another and/or relative to the continuum's endpoints.

In the literature on good measurement in the social sciences, a number of truisms about effective questionnaire design have been widely accepted. For example, the words used in questions should be understandable to all respondents, and the meanings imputed to those words should be as universal as possible among respondents (e.g., Oppenheim, 1992; Warwick & Lininger, 1975). Second, question wordings should avoid bias that would push answers one way or another (e.g., Parten, 1950; Young, 1939). Third, in order to minimize the impact of the idiosyncrasies of item wordings, it is best to aggregate answers to a battery of items into a single index (e.g., Likert, 1932; Thurstone, 1928). And the items used should be the few most efficient and effective ones tapping the construct of interest, to maximize validity, minimize respondent burden, and minimize the financial costs of data collection.

During the past five decades, social scientists have been following this advice and building batteries of questions to measure many important constructs, and this book documents just how much has been accomplished by these individuals: a lot. Thanks to large-scale collective enterprises, such as the National Election Study, small-scale research projects by lone investigators, and everything in between, we now have batteries to measure ideology, political partisanship, trust in government, political alienation and efficacy, racial attitudes, international attitudes, political information, values, participation, and much more. Most of these batteries are tried and true, having been employed in many

Measures of Political Attitudes

empirical investigations that reassure us about their validity and reliability. And they have been built largely following the general pieces of advice about good questionnaire design mentioned above.

Beyond that very general advice, however, there has been relatively little empirically validated and widely accepted wisdom in the questionnaire design literature about exactly how to word and structure the individual items that compose a battery. Consequently designers of most of the batteries in this book were left to their own devices when making a series of necessary decisions. For example, when constructing closed-ended questions, 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?

Every researcher's goal is to maximize the reliability and validity of the data he or she collects, so each of these design decisions should presumably be made so as to maximize these two indicators of data quality. Fortunately thousands of empirical studies provide clear and surprisingly unanimous advice on the issues listed above, but most of these studies have been unacknowledged in contemporary reviews of this literature (see, e.g., Bradburn, Sudman, & Associates, 1981; Converse & Presser, 1986; Schuman & Presser, 1981; Sudman, Bradburn, & Schwarz, 1996). Consequently it should come as no surprise that many batteries described in this book do not conform to the guidelines for good measurement suggested by this literature.

This chapter previews portions of a forthcoming book that will review this literature in detail, bringing to bear a rich set of evidence dating from the beginning of the 20th Century to the present (Krosnick & Fabrigar, in press). Most of this evidence comes from experimental studies comparing one method of question construction to another. And when brought together, these studies are remarkably consistent with one another in suggesting clear guidelines about how to maximize reliability and validity. This chapter offers a brief review of some of the implications of this literature, particularly in light of the sorts of items employed in the chapters that follow. The chapter's primary goal is to help readers see the design strengths of some batteries presented in later chapters and to suggest possibilities for experimentation with other batteries in order to increase their reliability and validity through subtle redesign.

The chapter begins with an issue of relevance to most chapters in the book: acquiescence. Subsequent sections consider issues of relevance to only some chapters: (1) the design of rating scales, including how many points should be offered, verbal labeling of them, and the branching approach to asking multiple, interrelated questions; (2) the impact of the order of response choices on answers; (3) the potential value of offering "no opinion" options to respondents and of directly measuring dimensions of attitude strength; (4) the impact of social desirability response bias on data quality; (5) the relative merits of rating scales versus ranking tasks; (6) the relative merits of open-ended versus closed-ended questions; and (7) the validity of questions asking people to describe their own mental processes.

Acquiescence

Defining Acquiescence

Many items in this book offer response choices such as "agree or disagree," "true or false," or "yes or no" (see especially Chapters 3–9). This sort of item format is very appealing from a practical standpoint because such items are easy to write. If one wants to identify

people who have positive attitudes toward bananas, for example, one simply needs to write a statement expressing an attitude (e.g., "I like bananas") and ask people whether they agree or disagree with it or whether it is true or false. Also, these formats can be used to measure a wide range of different constructs efficiently. Instead of having to change the response options from one question to the next as one moves from measuring liking to frequency to probability, one can use the same set of response options without having to reread them to respondents. In line with this logic, a series of studies suggests that it takes people about 75% longer to answer a multiple choice test question than to answer a comparable true-false question (e.g., Wesman, 1947). The popularity of agree–disagree and true–false item formats is, therefore, no surprise.

Despite this popularity, there has been a great deal of concern expressed over the years that these question formats may be seriously problematic. The danger is that some respondents may sometimes say "agree," "true," or "yes," regardless of the question being asked of them. So, for example, a person might agree with a statement that the United States should forbid speeches against democracy and might also agree with a statement that the United States should allow such speeches. This behavior, labeled *acquiescence,* can be defined as endorsement of an assertion made in a question, regardless of the content of the assertion.

In theory, acquiescence could result from a desire to be polite rather than confrontational in interpersonal interactions (Leech, 1983), from a desire of individuals of lower social status to defer to individuals of higher social status (Lenski & Leggett, 1960), or from an inclination to satisfice rather than optimize when answering questionnaires (Krosnick, 1991). According to this latter explanation, people sometimes shortcut the cognitive processes they execute when answering questions, and when they do so, they fall prey to a confirmatory bias. This bias inclines people toward accepting assertions, rather than thinking more extensively and seeing the flaws in those assertions.

Documenting Acquiescence

The evidence documenting acquiescence is now voluminous and consistently compelling, based on a range of different demonstration methods. For example, consider first just agree–disagree questions. When people are given such answer choices, are not told any questions, and are asked to guess what answers an experimenter is imagining, people guess "agree" much more often than "disagree" (e.g., Berg & Rapaport, 1954). In other studies pairs of statements were constructed stating mutually exclusive views (e.g., "I enjoy socializing" versus "I don't enjoy socializing"), and people were asked to agree or disagree with both. Although answers to such pairs should be strongly negatively correlated, 41 studies yielded an average correlation of only $-.22$. This correlation may be far from -1.0 partly because of random measurement error, but studies that corrected for such error suggest that the departure from -1.0 is also because of acquiescence.

Consistent with this claim, combining across 10 studies, an average of 52% of people agreed with an assertion, whereas an average of only 42% of people disagreed with the opposite assertion. Thus, people are apparently inclined toward agreeing rather than disagreeing, manifesting what might be considered an acquiescence effect of 10%. Another set of 8 studies compared answers to agree–disagree questions with answers to forced choice questions where the order of the views expressed by the response alternatives was the same as in the agree–disagree questions. An average of 14% more people agreed with an assertion than expressed the same view in the corresponding forced-choice question. Averaging across 7 studies, 22% of people, on average, agreed with both a statement and

its reversal, whereas only 10% of people disagreed with both. Thus, all of these methods suggest an average acquiescence effect of about 10%.

Other evidence indicates that acquiescence reflects a general tendency of some individuals across questions. For example, the average cross-sectional reliability of the tendency to agree with assertions is .65 across 29 studies. Furthermore, the over-time consistency of the tendency to acquiesce is about .75 over one month, .67 over four months, and .35 over four years (e.g., Couch & Keniston, 1960; Hoffman, 1960; Newcomb, 1943).

These same sorts of results (regarding correlations between opposite assertions, endorsement rates of items, their reversals, forced-choice versions, and so on) have been produced in studies of true–false questions and of yes–no questions, suggesting that acquiescence is present in these items as well. And there is other evidence regarding these response alternatives as well. For example, people are much more likely to answer yes–no factual questions correctly when the correct answer is "yes" than when it is "no" (e.g., Larkins & Shaver, 1967; Rothenberg, 1969), presumably because people are biased toward saying "yes." Similarly, a person's answer to a factual yes–no question (e.g., "Did you go shopping for food last week?") is more likely to disagree with an informant's account of that fact when the yes–no question is answered "yes" than when it is answered "no," again presumably because of a bias toward "yes" answers (Sigelman & Budd, 1986). When people say they are guessing the answer to a true–false question, 71% of answers were "true" in one study, and only 29% were "false." Acquiescence appeared just as clearly in studies using dichotomous items (e.g., "agree or disagree") and in studies offering more elaborate scales (e.g., "agree a lot, agree somewhat, neither agree nor disagree, disagree somewhat, disagree a lot").

The only body of evidence inconsistent with the acquiescence hypothesis involves forbid–allow questions. Rugg (1941) was the first to demonstrate that more people say "no" when asked whether something should be "allowed" than say "yes" when asked whether the same thing should be "forbidden" (see also Budd, Sigelman, & Sigelman, 1981; Schuman & Presser, 1981; Shaw & Budd, 1982). This pattern is in the opposite direction to what acquiescence would produce. Hippler and Schwarz (1986) argued that this tendency occurs because large numbers of people do not wish to take sides on these issues and "no" is the only response option offered that allows them to avoid taking a side. Such a bias toward saying "no" may have been quite a bit more common than acquiescence in responses to these items, masking it completely. Regardless, though, this is the only exception in the results of over 100 studies documenting the presence of acquiescence.

When Acquiescence Occurs

Acquiescence is most common among respondents of lower social status (e.g., Gove & Geerken, 1977; Lenski & Leggett, 1960), with less formal education (e.g., Ayidiya & McClendon, 1990; Narayan & Krosnick, 1996), of lower intelligence (e.g., Forehand, 1962; Hanley, 1959; Krosnick, Narayan, & Smith, 1996), of lower cognitive energy (Jackson, 1959), who do not like to think (Messick & Frederiksen, 1958), and of lower bias toward conveying a socially desirable image of themselves (e.g., Goldsmith, 1987; Shaffer, 1963). Also, acquiescence is most common when a question is difficult to answer (Gage, Leavitt, & Stone, 1957; Hanley, 1962; Trott & Jackson, 1967), after respondents have become fatigued answering a lot of prior questions (e.g., Clancy & Wachsler, 1971), and during telephone interviews rather than during face-to-face interviews (e.g., Calsyn, Roades, & Calsyn, 1992). Although some of these results are consistent with the notion that acquiescence results from politeness or deferral to people of higher social status, all of the results are

consistent with the satisficing explanation. Thus it appears that acquiescence occurs when people lack the skills and motivation to answer thoughtfully and when a question demands difficult cognitive tasks be executed in order for a person to answer precisely (Krosnick, 1991).

Correcting for Acquiescence

A number of studies now demonstrate how acquiescence can distort the substantive conclusions a researcher reaches from a study involving agree–disagree, true–false, or yes–no questions (e.g., Jackman, 1973; Winkles, Kanouse, & Ware, 1982). One of the best-known illustrations of the problem involved the F-scale used by Adorno and colleagues in *The Authoritarian Personality* to measure working-class authoritarianism (Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950). In fact, this scale turned out mostly to measure acquiescence (Christie 1991), and the substantive value of Adorno *et al.*'s. (1950) findings was completely undercut.

Although a number of methods for eliminating the distorting impact of acquiescence have been considered and employed over the years, only one seems to work effectively: avoiding agree–disagree, true–false, and yes–no question formats altogether. In order to understand why this is the optimal approach, it is useful to review the logic of other approaches and empirical evidence on their effectiveness.

One alternative method is based upon the presumption that certain people have acquiescent personalities and are likely to do all of the acquiescing. Therefore one simply needs to identify those people and statistically adjust their answers to correct for this tendency (e.g., Couch & Keniston, 1960). To this end many batteries of items have been developed to measure a person's tendency to acquiesce, and people who offer lots of "agree," "true," or "yes" answers across a large set of items can then be spotlighted as likely acquiescers. However, the evidence reviewed earlier suggests that acquiescence is not simply the result of having an acquiescent personality; rather, it is influenced by circumstantial factors as well. Because this "correction" approach does not take that into account, the corrections performed are not likely to fully and precisely account for acquiescence. Furthermore, if a set of agree–disagree items is asked of people and the acquiescers are then identified, there is no way to know how these people would have answered the questions had they not acquiesced. So post hoc statistical controlling for acquiescence seems unlikely to be fully effective.

Another popular technique thought to control acquiescence is measuring a construct with a large set of agree–disagree or true–false items, half of them making assertions opposite to the other half (called *item reversals;* see Altemeyer, 1996; Paulhus, 1991). This approach is designed to place acquiescers in the middle of the latent measurement dimension. However, it will do so only if the assertions made in the reversals are equally as extreme as the statements in the original items. Making sure this is true requires 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).

Even after all this effort, the balancing approach only partially solves the problem. Acquiescers who agree with every statement end up at a point on the measurement dimension where most of them probably do not belong. Instead, these people are arbitrarily placed there. And people who acquiesce on only some items end up with final scores that are closer to the dimension's midpoint than would validly represent their opinions. If

enough people acquiesce, arbitrarily placing the acquiescers (who have various distinctive characteristics, e.g., low education) at or near the dimension's midpoint can significantly distort correlations. If acquiescers were instead induced to answer items thoughtfully, their final index scores would be more valid than placing them at or near the midpoint. Nothing valid is learned from these people simply by balancing a battery; instead, valuable information about them is foregone.

The fatal flaw inherent in agree–disagree, true–false, and yes–no questions becomes obvious when one recognizes that answering such a question always requires a respondent to answer a comparable rating question in his or her mind first. For example, if a man is asked to agree or disagree with the assertion, "I am not a friendly person," he must first decide how friendly a person he is (perhaps concluding "extremely friendly"). Then he must translate that conclusion into the appropriate selection in order to answer the question he was asked ("disagree" to the original item). Researchers who use questions like this presume that the arraying of respondents along the agree–disagree dimension corresponds monotonically to the arraying of those individuals along the underlying substantive dimension of interest. That is, the more a person agrees with the assertion "I am not a friendly person," the lower he or she truly is in actual friendliness.

But consider the following scenario. Our hypothetical extremely friendly respondent answers a series of agree–disagree questions with stems such as "I am an extremely generous person," "I am never helpful to others," and "I always do well at everything I do." And the next stem in the question sequence is: "I am a friendly person." Given how extremely the previous stems were phrased, this one seems quite moderate (i.e., simply "friendly" instead of "extremely friendly"). Therefore our hypothetical respondent may feel that the word *friendly* does not adequately express the full extent of his gregariousness, so he may respond "disagree." Thus some people who disagree may feel they genuinely are not friendly, and other people who disagree may feel they are substantially more affable than the word *friendly* suggests. This clearly violates the monotonic equivalence of the response dimension and the underlying construct of interest.

This example points us to the solution to the acquiescence problem: Simply ask respondents directly how friendly they are. In fact, every agree–disagree, true–false, or yes–no question implicitly requires the respondent to rate an object along a continuous dimension in his or her mind, so asking about that dimension directly is bound to be less burdensome. In this light, it should be no surprise that the reliability and validity of rating-scale and forced-choice questions that present multiple competing points of view are higher than the reliability and validity of comparable agree–disagree, true–false, and yes–no questions, which focus on only a single point of view (e.g., Ebel, 1982; Mirowsky & Ross, 1991; Ruch & DeGraff, 1926; Wesman, 1946). Consequently it seems best to avoid agree–disagree, true–false, and yes–no formats altogether and instead ask just a few questions using other rating-scale or forced-choice formats.

Rating-Scale Formats

When designing a rating scale, one must begin by specifying the number of points on the scale, and the scales described in this book are of varying lengths, ranging from dichotomous items up to scales of 101 points (see, e.g., Chapters 11 and 12). A great number of studies have compared the reliability and validity of scales of varying lengths (for a review, see Krosnick & Fabrigar, in press). For bipolar scales, which have a neutral or status-quo point in the middle (e.g., running from positive to negative), reliability and validity are highest for about 7 points (e.g., Matell & Jacoby, 1971). In contrast, the reliabil-

ity and validity of unipolar scales, with a zero point at one end (e.g., running from no importance to very high importance), seem to be optimized for a bit shorter scales, approximately 5 points long (e.g., Wikman & Warneryd, 1990). Techniques such as magnitude scaling (e.g., Lodge, 1981), which offer scales with an infinite number of points, yield data of lower quality than do more conventional rating scales and should, therefore, be avoided (e.g., Cooper & Clare, 1981; Miethe, 1985; Patrick, Bush, & Chen, 1973).

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

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 response choices are presented visually, as in self-administered questionnaires, people are inclined toward selecting answer choices offered early in a list, yielding primacy effects (e.g., Krosnick & Alwin, 1987; Sudman, Bradburn, & Schwarz, 1996). But when the answer choices are read aloud to people, recency effects tend to appear, whereby people are inclined to select the options offered last (e.g., McClendon, 1991). These effects are most pronounced among respondents who have limited cognitive skills and when questions are more cognitively demanding (Krosnick & Alwin, 1987; Payne, 1949–1950). All this is consistent with the theory of satisficing (Krosnick, 1991), 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. Almost none of the scales described in this book are accompanied by advice to rotate response order, but the evidence of such effects is so consistent as to suggest that doing so is well worthwhile.

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., Converse, 1964), questionnaire designers have often explicitly offered respondents the option to say they have no opinion. And indeed, many more people say they "don't know" what their opinion is when this is done than when it is not (e.g., Schuman & Presser, 1981). People tend to offer this response under conditions that seem sensible (e.g., when they lack knowledge on the issue, Donovan & Leivers, 1993), and people prefer to be given this option in questionnaires (Ehrlich, 1964). Furthermore offering a "don't know" option significantly reduces the number of people who offer substantive evaluations of obscure or fictitious attitude objects, such as the Agricultural Trade Act (Schuman & Presser, 1981).

However, most "don't know" responses are due to conflicting feelings or beliefs (rather than lack of feelings or beliefs) or to uncertainty about exactly what a question's response alternatives mean or what the question is asking (e.g., Coombs & Coombs, 1976–1977). When people are pushed to offer an opinion instead of saying "don't know," the reliability and validity of data collected is no lower than when a "no opinion" option is offered and people are encouraged to select it (e.g., McClendon & Alwin, 1993). That is, people who would have selected this option if offered nonetheless report meaningful opinions when it is not offered. In fact, this is true even of opinions about obscure or fictitious attitude objects; people base their attitude reports on their best guesses about the objects' likely characteristics (Schuman & Presser, 1981). Therefore it is wise that most of the items described in this book do not offer explicit "don't know" response alternatives.

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) recently 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 seems preferable to ask people to describe where their attitude falls along the strength continuum.

Unfortunately there are many different aspects of attitude strength, and they are largely independent of each other (see, e.g., Krosnick, Boninger, Chuang, Berent, & Carnot, 1993). For example, people can be asked how important the issue is to them personally or how much they have thought about it or how certain they are of their opinion or how knowledgeable they feel about it (for details on measuring these and many other dimensions, see Wegener, Downing, Krosnick, & Petty, 1995). And one can measure the length of time it takes a person to answer a question, which reflects construct accessibility (Bassili & Fletcher, 1991). Each of these dimensions can help to differentiate attitudes that are crystallized and consequential from those that are not.

Social Desirability

The Notion of Social Desirability Response Bias

An issue potentially applicable to many chapters in this volume is that of social desirability response bias (for an earlier review, see Paulhus, 1991). One instance illustrating this potential problem in the context of politics involved the 1989 Virginia gubernatorial race, in which Douglas Wilder, an African-American Democrat, ran against Marshall Coleman, a Caucasian Republican. Preelection polls consistently gave Wilder a lead of between 4% and 11%, but on election day Wilder won by a mere .6%. Finkel, Guterbock, and Borg (1991) claimed the poll error was due to respondent dishonesty based on race. They argued that especially when interviewed by African Americans, Caucasian respondents are reluctant to express an intention to vote for a Caucasian candidate in a race against an African-American candidate. Finkel *et al.* (1991) demonstrated that a race-of-interviewer effect along these lines was especially powerful among respondents who did not identify with a major political party (i.e., "independents") and respondents who initially said they were undecided and were then pushed by the interviewer to express a candidate preference.

Finkel *et al.* (1991) concluded that this is evidence of a bias in reports toward presenting a socially desirable self-image to one's interviewer.

For this bias to have fully explained the error in all preelection polls in Virginia, a very large proportion of the interviewers involved would have had to be African-American and would have had to be identifiable as such by voice. This seems a bit implausible. Furthermore, there is another compelling possible explanation for the polls' error that does not involve race at all. In an extensive analysis of numerous candidate elections and referenda, Visser, Krosnick, Marquette, and Curtin (in press) showed that preelection polls have routinely overestimated the margin of victory of the winner. One possible explanation for this is the sort of bandwagon effect described by Noelle-Neumann's (1984) notion of the "spiral of silence." From publicity of preelection poll results, people often learn that a particular candidate is leading over his or her challengers. And when later interviewed for another poll, some supporters of a challenger may be reluctant to seem out of step with the majority and may, therefore, express support for the leading candidate or may say "don't know." This, too, would constitute measurement error due to intentional misrepresentation by a respondent to present a favorable self-image to an interviewer.

The Plausibility of the Threat

A number of different lines of research endorse the plausibility of the notion that people might lie to interviewers and researchers. For example, DePaulo, Kashy, Kirkendol, Wyer, and Epstein (1996) had people complete daily diaries in which they recorded any lies that they told during a seven-day period. On average, people reported telling one lie per day, with some people telling many more, and 91% of the lies involved misrepresenting oneself in some way. This evidence is in line with theoretical accounts from sociology (Goffman, 1959) and psychology (Schlenker & Weigold, 1989) asserting that an inherent element of social interaction is constructing an image of oneself in the eyes of others in pursuit of relevant goals. The fact that being viewed favorably by others is more likely to bring rewards and minimize punishments than being viewed unfavorably may motivate people to construct favorable images, sometimes via deceit. If this sort of behavior is common in daily life, why wouldn't people lie when answering questionnaires as well?

In fact, there are a number of reasons to believe that the motivation to lie in surveys might be minimal. First, when filling out an anonymous questionnaire, no rewards or punishments can possibly be at stake. And second, in most surveys and laboratory experiments, the respondent's relationships with an interviewer and/or a researcher are likely to be so short-lived and superficial that very little of consequence is at stake as well. Certainly even a small frown of disapproval from a total stranger can cause a bit of discomfort, but this is not likely to be especially noxious. And the cognitive task of figuring out which response to each question will garner the most respect from an interviewer and/or a researcher is likely to be demanding enough to be worth doing only when the stakes are significant. So perhaps there is not so much danger here after all. And perhaps the tendency of preelection polls to overpredict the margin of victory of the winner is due to a process having nothing to do with intentional lying to present a respectable image of oneself to others.

It would be nice if that were true, but unfortunately there is another potential source of systematic distortion in responses to even self-administered anonymous questionnaires: self-deception. Not only do people want to maintain favorable images of themselves in the eyes of others, but they want to have such images in their own eyes as well. According to many psychological analyses, the pursuit of self-esteem is a basic human motive (see, e.g.,

Sedikides & Strube, 1997), and it is driven partly by such inevitable realities as the prospect of death (e.g., Greenberg, Solomon, & Pyszczynski, 1997). So people may be motivated to convince themselves that they are respectable, good people, and doing so may at times entail misconstrual of facts (see Paulhus, 1984, 1986, 1991). If people fool themselves in this way, such misconstrual will find its way into questionnaire responses, even when respondents want to accurately report their perceptions to an interviewer and/or a researcher. Obviously it is tricky business to fool oneself because part of the mind might need to know that it is fooling another part. But such self-deception can be so automatic and may even unfold outside of consciousness that people would not be aware of it at all.

What a mess! Questionnaire research is based on the assumption that people can and will accurately report information. If this is not true, either because of other-deception or because of self-deception, this threatens the value of questionnaire-based data. Many of the scales described in this book are potentially vulnerable to this problem because it seems socially desirable to express interest and involvement in politics, not to express racial prejudice, to endorse classically American values instead of repudiating them, to appear nationalistic instead of cynical about one's own country, and so on. Given this threat, researchers have been very interested over the years in exploring whether social desirability response bias is truly a source of data distortion, assessing its magnitude, and developing techniques to overcome it.

Evidence of Social Desirability Bias

The evidence documenting systematic and intentional misrepresentation is now quite voluminous and very convincing, partly because the same conclusion has been supported by studies using many different methods. One such method is the *bogus pipeline technique,* which involves telling respondents that the researcher can otherwise determine the correct answer to a question they will be asked, so they might as well answer it accurately (see, e.g., Roese & Jamieson, 1993). Under these conditions, people are more willing to report substance use (Evans, Hansen, & Mittlemark, 1977; Murray & Perry, 1987). Likewise, white respondents are more willing to ascribe undesirable personality characteristics to African-Americans (Pavlos, 1972, 1973; Sigall & Page, 1971) and are more willing to report disliking African Americans (e.g., Allen, 1975) under bogus pipeline conditions. Women are less likely to report supporting the women's movement under bogus pipeline conditions than under normal reporting conditions (Hough & Allen, 1975). And people are more likely to admit having been given secret information under bogus pipeline conditions (Quigley-Fernandez & Tedeschi, 1978).

Another approach to documenting such distortion is to compare responses given when people believe their answers will have significant consequences for them to responses given when no such consequences exist. For example, in one study respondents who believed that they had already been admitted to an apprenticeship program admitted to having less respectable personality characteristics than did comparable respondents who believed they were being evaluated for possible admission to the program (Michaelis & Eysenck, 1971).

Yet another approach to this problem involves the *randomized response technique* (Warner, 1965). Here respondents answer one of various different questions, depending on what a randomizing device instructs. Thus the interviewer and the researcher do not know exactly which question each person is answering, so the respondents can presumably feel freer to be honest. In one such study Himmelfarb and Lickteig (1982) had respondents secretly toss three coins before answering a yes–no question. Respondents were instructed to say "yes" if all three coins came up heads, to say "no" if all three coins came up tails,

and to answer the yes–no question truthfully if any combination of heads and tails came up. People answering in this fashion admitted to falsifying their income tax reports and enjoying soft-core pornography more than did respondents who were asked these questions directly.

Still another approach to assessing the impact of social desirability is by studying interviewer effects. The presumption here is that the observable characteristics of an interviewer may suggest to a respondent which answers he or she would consider most respectable. So if answers vary in a way that corresponds with interviewer characteristics, it suggests that respondents tailored their answers accordingly. For example, various studies have found that African Americans report more favorable attitudes toward whites when their interviewer is white than when the interviewer is African-American (Anderson, Silver, & Abramson, 1988a, 1988b; Campbell, 1981; Schuman & Converse, 1971). Likewise, white respondents express more favorable attitudes toward African Americans and the principle of racial integration to African-American interviewers than to white interviewers (Campbell, 1981; Cotter, Cohen, & Coulter, 1982; Finkel *et al.*, 1991). In another study, people expressed more positive attitudes toward firefighters when they thought their interviewer was a firefighter than when they did not hold this belief (Atkin & Chaffee, 1972–1973).

Another approach to this issue involves comparisons of different modes of data collection. In general, pressure to appear socially desirable is presumably greatest when a respondent is being interviewed by another person, either face-to-face or over the telephone. But when respondents complete written questionnaires alone, this pressure is presumably lessened. Consistent with this reasoning, Catholics in one study were more likely to report favoring legalized abortion and birth control when completing self-administered questionnaires than when being interviewed by telephone or face-to-face (Wiseman, 1972). And people report being happier with their lives in interviews than on self-administered questionnaires (Cheng, 1988).

The anonymity of some self-administered questionnaires further reduces social pressure, so it, too, offers an empirical handle for addressing this issue. In one study, Gordon (1987) asked respondents about dental hygiene on questionnaires; half the respondents (selected randomly) were asked to write their names on the questionnaires, whereas the other half were not. Dental checkups, brushing, and flossing were all reported to have been done more often when people wrote their names on the questionnaires than when they did not. Thus socially desirable responses were apparently more common under conditions of high identifiability. Likewise people reported having more desirable personality characteristics when they wrote their names, addresses, and telephone numbers on questionnaires than when they did not (Paulhus, 1984).

Taken together, these studies all suggest that some people sometimes distort their answers in surveys in order to present themselves as having more socially desirable or respectable characteristics or behavioral histories. But the social desirability driven distortions documented above represent only those involving other deception. There may be significant amounts of self-deception going on as well, and when combined with other deception, social desirability driven error may be even more substantial. Needless to say, there is no easy way of documenting self-deception in studies that involve only self-reports. When records can be checked, researchers can validate self-reports against, for example, official records of whether a person voted in a particular election. So if respondents are asked (using a randomized response technique, for example) whether they voted or not in an election and more people say they voted than can be confirmed in the official records, that would provide a suggestive estimation of the magnitude of self-deception. Unfortunately, few, if any, such studies have been

done, so it is difficult to draw any conclusions at the moment about the prevalence of such error in data.

Controlling for Social Desirability Response Bias

One approach to correcting for social desirability bias involves making a big assumption. According to this perspective, some people are especially likely to distort their responses to all questions in socially desirable directions, whereas other people are especially unlikely to do so. Therefore all respondents can be asked a battery of questions with strong social desirability connotations, and the people who offer especially large numbers of socially desirable responses would be considered suspect. A researcher could then retest hypotheses after removing these respondents from a data set to see what effect doing so has. Also a researcher can statistically control for the tendency to answer such a battery in a socially desirable fashion when analyzing correlational associations among other variables in a data set (see Paulhus, 1991).

The big assumption involved in this approach is that the tendency to answer one set of questions with a social desirability bias can effectively predict the extent of such bias in a single other question. But it seems likely that whether a particular person answers this single other question with such bias depends on a number of other factors (e.g., the match of the race of the respondent and the interviewer if the question involves a racial issue, but not if the topic is unrelated to race). Therefore this approach probably at best can detect only a portion of the social desirability bias present and may in fact claim to detect such bias in answers that are not in fact thusly contaminated.

Perhaps the most important reason to hesitate when considering this correction method is that it involves after-the-fact adjustment. Methods like the bogus pipeline and the randomized response technique lead each respondent to answer honestly, so valid data are available for all respondents. But when suspect respondents are identified by their answers to a social desirability battery, there is no way to know how they would have answered target questions if they had done so accurately. Statistical correction using answers to a battery is also suspect, as there is no way to know whether a bias toward socially desirable answers among suspect respondents had no real impact because the "correct" answers to target questions for these people also happened to be the socially desirable ones. Thus there is good reason to hesitate before presuming that statistical detection or correction using social desirability response bias batteries is an effective solution to the problem.

The better approach to solving the intentional misrepresentation problem is employing one or more of the techniques outlined above that either reduce pressure to appear socially desirable (e.g., via anonymity) or create new pressures to be honest (e.g., via the bogus pipeline). If a researcher is concerned that social desirability pressures might be distorting answers to a particular question, that hypothesis can be evaluated in a pretest by randomly assigning some respondents to answer using an ordinary self-report approach and other respondents to provide reports in a way that reduces the likelihood of social desirability based distortion. If the distribution of answers is different in these two groups, then one knows that a measurement method must be employed in the final study that solves this problem (e.g., by allowing completely anonymous responses).

Promising New Techniques

Very new techniques are being developed to solve not only the intentional misrepresentation problem but also the self-deception problem. One such technique involves bypassing respondents' reports altogether and measuring cognitive processes more directly. For example, people's attitudes toward an object are revealed by tiny movements of fac

muscles upon observation of the object (Cacioppo, Petty, Losch, & Kim, 1986), and electrical activity in the brain indicates attitudes as well (Crites, Cacioppo, Gardner, & Berntson, 1995). Measurement of such phenomena is quite cumbersome and not suitable for most surveys and laboratory experiments, but another measure—reaction time—may be more practical.

According to an accumulating body of studies, the length of time it takes a person to make a judgment can be used to gauge evaluations. For example, a person may be asked to place his or her right index finger on one key of a computer keyboard and his or her left index finger on another key. Then he or she may be asked to read a word that appears in the middle of the computer screen and decide as quickly as possible whether it refers to a good–pleasant concept or a bad–unpleasant concept. The person can be instructed to press the right-hand button in the former case and the left-hand button in the latter case. The computer can then measure the amount of time between the appearance of a word, such as "good" or "bad," and a button press.

Recent research has shown that the length of time it takes to press the button, usually a fraction of a second, can be influenced by a very fast, subliminal flash of another word on the computer screen, just before the appearance of the word to be evaluated (e.g., Hermans, De Houwer, & Eelen, 1994). If the subliminal flash is of a positive word (e.g., "nice"), people are a little bit quicker at recognizing that "good" is a positive word. If the subliminal word is negative (e.g., "rotten"), then people are a little bit slower at recognizing that "good" is a positive word. This technique can be used to measure people's attitudes toward objects when they think they are simply being asked to identify words as good and bad.

If on a given trial the subliminal word is "pizza" and the supraliminal word is "good," then a researcher can assess whether pizza speeds up or slows down identification of "good" as a positive word and by how much. The more speeding up a person manifests, the more positive his or her attitude toward pizza presumably is. And the more slowing down a person manifests, the more negative his or her attitude toward pizza probably is. Fazio, Jackson, Dunton, and Williams (1995) have used this technique to measure racial prejudice and have found evidence that such measurements are quite valid. In theory, these attitude measurements may be more valid than self-reports because the former are uncontaminated by any other-deception or self-deception biases. However, some new evidence suggests that self-reports may be more accurate predictors of some behaviors, whereas the reaction-time measurements may be better predictors of other behaviors (e.g., Dovidio, Kawakami, Johnson, Johnson, & Howard, 1997).

This procedure is obviously easy to execute in a laboratory, and it is also easy to execute in in-home surveys when field interviewers carry laptop computers. Although some recent work suggests that reaction-time measurement can be done validly over the telephone (e.g., Bassili & Fletcher, 1991), no evidence yet documents whether the very quick reaction-time differences of interest here can be captured with that methodology. Perhaps more importantly, ethical considerations are raised by the fact that respondents are asked to participate in a procedure but are not told what is being measured. Nonetheless, there may be some potential use for this technique in survey research generally, and it may provide a way to overcome social desirability-based response distortion.

Rating versus Ranking

Although most questions in this book involve rating scales, some involve ranking tasks (see, e.g., Chapter 11). In fact, many have argued that ranking is the superior method for measuring political values (e.g., Ingelhart, 1977; Rokeach, 1973). And the choice

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

Surprisingly, however, rankings are more effective than ratings because ratings suffer from a significant problem: non-differentiation. According to Krosnick's (1991) theory of survey satisficing, some respondents are not especially motivated to think carefully about the questions they are asked, or doing the cognitive work required is especially demanding for them. Under such circumstances, people may choose to shortcut the response process via a series of specific response strategies. When a person is asked to rate a large set of objects on a single scale, one satisficing strategy is to select what appears to be a reasonable point to rate most objects on the scale and rate all objects at or near that point (i.e., non-differentiation), rather than thinking carefully about each object and rating different objects differently (see Krosnick, 1991; Krosnick & Alwin, 1988; Krosnick, Narayan, & Smith, 1996). So, for example, if a respondent is asked to rate the importance of a series of objects (e.g., child qualities such as honesty, intelligence, and responsibility), satisficers may rate each one "very important" (Krosnick & Alwin, 1988). At least partly as a result, the reliability and validity of ranking data are superior to those of rating data (e.g., Miethe, 1985; Munson & McIntyre, 1979; Nathan & Alexander, 1985; Rankin & Grube, 1980; Reynolds & Jolly, 1980).

Batteries of items using the same response scale (as many of the batteries in this book do) are at risk for this sort of satisficing. Such non-differentiation is most likely to occur when respondent motivation and/or ability to answer optimally are low (e.g., Krosnick, Narayan, & Smith, 1996). Therefore, the problem might be reduced by taking steps to minimize the cognitive difficulty of the items and to maximize respondent motivation to answer carefully and precisely (see Krosnick, 1991). However, it is not yet clear whether such strategies can be completely effective. Therefore, although rankings do not yield interval-level measures of the perceived distances between objects in respondents' minds and are more statistically cumbersome to analyze (see Alwin & Jackson, 1982), these measures are apparently more useful when a researcher's goal is to ascertain rank orders of objects. At the very least, researchers interested in implementing a series of ratings might consider ordering questions so that no two adjacent questions offer the same response alternatives.

Open versus Closed Questions

Although the vast majority of items listed in this book are closed-ended (meaning that they ask respondents to choose among offered sets of response choices), a few are open-ended, allowing respondents to answer in their own words (see, e.g., Chapters 10, 11, and 12). For example, one of the most frequently asked and widely publicized survey items inquires about what people consider to be the most important problem facing the country, usually presented in an open-ended format (see Chapter 11). But a closed-ended version of this

question might be used instead, asking, "What is the most important problem 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 1000 respondents, nearly 1000 different answers will be given to the "most important problem" question if considered word for word. But in order to analyze these answers, they must be clumped into a relatively small number of categories. This requires that researchers must develop a coding scheme for each question; multiple people must read and code the answers into the categories; the level of agreement between the coders must be ascertained; and the procedure must be refined and repeated if agreement is too low. The time and financial costs of such a procedure, coupled with the added challenge of requiring interviewers to carefully transcribe answers, have led many researchers to favor closed-ended questions, which in essence ask respondents to directly code themselves into categories that the researcher specifies.

When closed-ended questions are used to ascertain categorical judgments of this sort (where the options represent different objects rather than different points along a single continuum), researchers often do not want to confine respondents to the list, so they offer an "other" response alternative. However, respondents tend to confine their answers to the choices offered, even when the "other" opportunity is offered (Jenkins, 1935; Lindzey & Guest, 1951). If the list of choices presented by a question is incomplete, even the rank ordering of the choices that are explicitly offered can be different from what would be obtained if a longer or shorter list were offered instead. Therefore a closed-ended categorical question can be used effectively only if its answer choices are comprehensive, and this can usually be assured only if an open-ended version of the question is administered in a pretest using a reasonably large sample. Given that, it may be more practical simply to ask the open-ended question in the final survey.

Introspection

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. For example, political scientists have routinely asked people to explain why they voted as they did in a particular election (see Chapter 11). Whether employing an open-ended question or a closed-ended one, this approach requires people to introspect and describe their own cognitive processes, which was one of modern psychology's first core research methods (Hothersall, 1984).

Early in this century, though, it became clear that this method can often yield misleading results (Hothersall, 1984). And Nisbett and Wilson (1977) articulated an argument about why this is so, reviewing a great deal of evidence in support of their theoretical account. Studies done since their landmark paper have further 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 viewed skeptically.

Conclusion

This review illustrates only a tiny fraction of the wealth of knowledge about questionnaire design buried in the journals of many social science disciplines. Many, if not most, of the

batteries described in this book were developed long before the accumulated wisdom of this literature was apparent. Consequently many of these batteries do not fully conform to the advice offered in this chapter. It clearly seems worthwhile, then, to consider experimenting in future studies with slight alterations in format to see whether reliability and validity can be improved.

Some readers, after plowing through the above review, might say instead: How much can we hope to gain from such efforts? Aren't the reliabilities and validities sufficiently high to suggest that the batteries are just fine as they are? Furthermore why should we view these items as potentially "broken" when they have been used successfully in numerous investigations? This is certainly an understandable perspective, so it seems quite reasonable that some readers might feel this way.

However, the literature backing the above advice is both so voluminous and so consistent in its findings that it is hard to disregard the possibility that the accuracy of just about any measuring instrument can be improved by following it. Likewise what may appear to be high reliability in some scales may actually be highly reliable systematic measurement error rather than reliable substantive assessment. And correlations between items that appear to suggest high validity may instead be associations due to systematic measurement error rather than substance (see, e.g., Krosnick & Alwin, 1988). So batteries that appear not to be "broken" may indeed be improvable through systematic, theory-guided experimentation. And the literature reviewed in this chapter points to some possible avenues for such exploration.

Acknowledgments

This chapter was written partly while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences, supported by National Science Foundation Grant SBR-9022192. The author wishes to express his thanks to Michael Tichy for his help with manuscript preparation. Correspondence should be addressed to Jon A. Krosnick, Department of Psychology, Ohio State University, 1885 Neil Avenue, Columbus, Ohio 43210 (e-mail: Krosnick@osu.edu).

References

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

Allen, B. P. (1975). Social distance and admiration reactions of "unprejudiced" whites. *Journal of Personality, 43,* 709–726.

Altemeyer, B. (1996). *The authoritarian specter.* Cambridge, MA: Harvard University Press.

Alwin, D. F., & Jackson, D. J. (1982). Adult values for children: An application of factor analysis to ranked preference data. In K. F. Schuessler (Ed.), *Sociological methodology 1980* (pp. 311–329). San Francisco: Jossey-Bass.

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

Anderson, B. A., Silver, B. D., & Abramson, P. R. (1988b). The effects of the race of the interviewer on race-related attitudes of black respondents in SRC/CPS national election studies. *Public Opinion Quarterly, 52,* 289–324.

Atkin, C. K., & Chaffee, S. H. (1972–1973). Instrumental response strategies in opinion interviews. *Public Opinion Quarterly, 36,* 69–79.

Ayidiya, S. A., & McClendon, M. J. (1990). Response effects in mail surveys. *Public Opinion Quarterly, 54,* 229–247.

Bassili, J. N., & Fletcher, J. F. (1991). Response-time measurement in survey research. *Public Opinion Quarterly, 55,* 331–346.

Berg, I. A., & Rapaport, G. M. (1954). Response bias in an unstructured questionnaire. *Journal of Psychology, 38,* 475–481.

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

Budd, E. C., Sigelman, C. K., & Sigelman, L. (1981). Exploring the outer limits of response bias. *Sociological Focus, 14,* 297–307.

Cacioppo, J. T., Petty, R. E., Losch, M. E., & Kim, H. S. (1986). Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology, 50,* 260–268.

Calsyn, R. J., Roades, L. A., & Calsyn, D. S. (1992). Acquiescence in needs assessment studies of the elderly. *Gerontologist, 32,* 246–252.

Campbell, B. A. (1981). Race-of-interviewer effects among southern adolescents. *Public Opinion Quarterly, 45,* 231–244.

Christie, R. (1991). Authoritarianism and related constructs. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 501–571). San Diego, CA: Academic Press.

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

Clancy, K. J., & Wachsler, R. A. (1971). Positional effects in shared-cost surveys. *Public Opinion Quarterly, 35,* 258–265.

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

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.

Coombs, C. H., & Coombs, L. C. (1976–1977). "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.

Cotter, P., Cohen, J., & Coulter, P. B. (1982). Race of interviewer effects in telephone interviews. *Public Opinion Quarterly, 46,* 278–294.

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

Crites, S. L., Jr., Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1995). Bioelectrical echoes from evaluative categorization: II. A late positive brain potential that varies as a function of attitude rather than attitude report. *Journal of Personality and Social Psychology, 68,* 997–1013.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70,* 979–995.

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.

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

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology, 33,* 510–540.

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,* 221–231.

Evans, R. I., Hansen, W. B., & Mittlemark, M. B. (1977). Increasing the validity of self-reports of smoking behavior in children. *Journal of Applied Psychology, 62,* 521–523.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013–1027.

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.

Forehand, G. A. (1962). Relationships among response sets and cognitive behaviors. *Educational and Psychological Measurement, 22,* 287–302.

Gage, N. L., Leavitt, G. S., & Stone, G. C. (1957). The psychological meaning of acquiescence set for authoritarianism. *Journal of Abnormal and Social Psychology, 55,* 98–103.

Goffman, E. (1959). *The presentation of self in everyday life.* Garden City, NY: Doubleday/Anchor Books.

Goldsmith, R. E. (1987). Two studies of yeasaying. *Psychological Reports, 60,* 239–244.

Gordon, R. A. (1987). Social desirability bias: A demonstration and technique for its reduction. *Teaching of Psychology, 14,* 40–42.

Gove, W. R., & Geerken, M. R. (1977). Response bias in surveys of mental health: An empirical investigation. *American Journal of Sociology, 82,* 1289–1317.

Greenberg, J., Solomon, S., & Pyszczynski, T. (1997). Terror management theory of self-esteem and cultural worldviews: Empirical assessments and conceptual refinements. *Advances in Experimental Social Psychology, 29,* 61–139.

Hanley, C. (1959). Responses to the wording of personality test items. *Journal of Consulting Psychology, 23,* 261–265.

Hanley, C. (1962). The "difficulty" of a personality inventory item. *Educational and Psychological Measurement, 22,* 577–584.

Hermans, D., De Houwer, J., & Eelen, P. (1994). The affective priming effect: Automatic activation of evaluative information in memory. *Cognition and Emotion, 8,* 515–533.

Himmelfarb, S., & Lickteig, C. (1982). Social desirability and the randomized response technique. *Journal of Personality and Social Psychology, 43,* 710–717.

Hippler, H.-J., & Schwarz, N. (1986). Not forbidding isn't allowing: The cognitive basis of the forbid-allow asymmetry. *Public Opinion Quarterly, 50,* 87–96.

Hoffman, P. J. (1960). Social acquiescence and "education." *Educational and Psychological Measurement, 20,* 769–776.

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

Hough, K. S., & Allen, B. P. (1975). Is the "women's movement" erasing the mark of oppression from the female psyche? *Journal of Psychology, 89,* 249–258.

Ingelhart, R. (1977). *The silent revolution.* Princeton, NJ: Princeton University Press.

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

Jackson, D. N. (1959). Cognitive energy level, acquiescence, and authoritarianism. *Journal of Social Psychology, 49,* 65–69.

Jackson, D. N. (1967). Acquiescence response styles: Problems of identification and control. In I. A. Berg (Ed.), *Response set in personality assessment* (pp. 71–114). Chicago: Aldine.

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

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

Krosnick, J. A. (1991). 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., & 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., & Fabrigar, L. R. (In press). *Designing good questionnaires: Insights from psychology.* New York: Oxford University Press.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation, 70,* 29–44.

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.

Larkins, A. G., & Shaver, J. P. (1967). Matched-pair scoring technique used on a first-grade yes-no type economics achievement test. *Utah Academy of Science, Art, and Letters: Proceedings, 44-I,* 229–242.

Leech, G. N. (1983). *Principles of pragmatics.* London: Longman.

Lenski, G. E., & Leggett, J. C. (1960). Caste, class, and deference in the research interview. *American Journal of Sociology, 65,* 463–467.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 104,* 44–53.

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

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

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.

Messick, S., & Frederiksen, N. (1958). Ability, acquiescence, and "authoritarianism." *Psychological Reports, 4,* 687–697.

Michaelis, W., & Eysenck, H. J. (1971). The determination of personality inventory factor patterns and intercorrelations by changes in real-life motivation. *Journal of Genetic Psychology, 118,* 223–234.

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

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

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.

Murray, D. M., & Perry, C. L. (1987). The measurement of substance use among adolescents: When is the bogus pipeline method needed? *Addictive Behaviors, 12,* 225–233.

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

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60,* 58–88.

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

Newcomb, T. E. (1943). *Personality and social change.* New York: Dryden Press.

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

Noelle-Neumann, E. (1984). *The spiral of silence.* Chicago: University of Chicago Press.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing, and attitude measurement.* London: Pinter Publishers.

Parten, M. (1950). *Surveys, polls, and samples: Practical procedures.* New York: Harper and Brothers.

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.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46,* 598–609.

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. Wiggins (Eds.), *Personality assessment via questionnaires: Current issues in theory and measurement* (pp. 143–165). New York: Springer-Verlag.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightman (Eds.), *Measures of personality and social psychological attitudes,* Vol. 1 (pp. 17–59). San Diego: Academic Press.

Pavlos, A. J. (1972). Racial attitude and stereotype change with bogus pipeline paradigm. *Proceedings of the 80th Annual Convention of the American Psychological Association, 7,* 292.

Pavlos, A. J. (1973). Acute self-esteem effects on racial attitudes measured by rating scale and bogus pipeline. *Proceedings of the 81st Annual Convention of the American Psychological Association, 8,* 165–166.

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

Quigley-Fernandez, B., & Tedeschi, J. T. (1978). The bogus pipeline as lie detector: Two validity studies. *Journal of Personality and Social Psychology, 36,* 247–256.

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.

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

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin, 114,* 363–375.

Rokeach, M. (1973). *The nature of human values.* New York: Free Press.

Rothenberg, B. B. (1969). Conservation of number among four- and five-year-old children: Some methodological considerations. *Child Development, 40,* 383–406.

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.

Rugg, D. (1941). Experiments in wording questions: II. *Public Opinion Quarterly, 5,* 91–92.

Schlenker, B. R., & Weigold, M. F. (1989). Goals and the self-identification process: Constructing desires identities. In L. A. Pervin (Ed.), *Goal concepts in personality and social psychology* (pp. 243–290). Hillsdale, NJ: Erlbaum.

Schuman, H., & Converse, J. M. (1971). The effect of black and white interviewers on black responses. *Public Opinion Quarterly, 35,* 44–68.

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

Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. *Advances in Experimental Social Psychology, 29,* 209–269.

Shaffer, J. W. (1963). A new acquiescence scale for the MMPI. *Journal of Clinical Psychology, 19,* 412–415.

Shaw, J. A., & Budd, E. C. (1982). Determinants of acquiescence and naysaying of mentally retarded persons. *American Journal of Mental Deficiency, 87,* 108–110.

Sigall, H., & Page, R. (1971). Current stereotypes: A little fading, a little faking. *Journal of Personality and Social Psychology, 18,* 247–255.

Sigelman, C. K., & Budd, E. C. (1986). Pictures as an aid in questioning mentally retarded persons. *Rehabilitation Counseling Bulletin, 29,* 173–181.

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.

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

Sussman, B. (1985, November 28). Hidden racial attitudes distorted Virginia polls. *The Washington Post.*

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

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33,* 529–554.

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

Trott, D. M., & Jackson, D. N. (1967). An experimental analysis of acquiescence. *Journal of Experimental Research in Personality, 2,* 278–288.

Visser, P. S., Krosnick, J. A., Marquette, J., & Curtin, M. (In press). Improving election forecasting: Allocation of undecided respondents, identification of likely voters, and response order effects. In P. Lavrakas & M. Traugott (Eds.), *Election polls, the news media, and democracy.*

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60,* 63–69.

Warwick, D. P., & Lininger, C. A. (1975). *The sample survey: Theory and practice.* New York: McGraw-Hill.

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

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.

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

Wesman, A. G. (1947). Active versus blank responses to multiple-choice items. *Journal of Educational Psychology, 38,* 89–95.

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

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

Young, P. V. (1939). *Scientific social surveys and research.* New York: Prentice-Hall.