Question and Questionnaire Design

Jon A. Krosnick

Stanford University

and

Stanley Presser

University of Maryland

February 15, 2009

To appear in the

Handbook of Survey Research (2nd Edition)
James D. Wright and Peter V. Marsden (Eds).
San Diego, CA: Elsevier.

Jon Krosnick is University Fellow at Resources for the Future. Address correspondence to Jon A. Krosnick, Stanford University, 432 McClatchy Hall, Stanford, CA 94305 (email: Krosnick@stanford.edu) or Stanley Presser, 4121 Art-Sociology Building, University of Maryland, College Park, MD 20742-1315 (email: spresser@socy.umd.edu).

The heart of a survey is its questionnaire. Drawing a sample, hiring and training interviewers and supervisors, programming computers, and other preparatory work is all in service of the conversation that takes place between researchers and respondents. Survey results depend crucially on the questionnaire that scripts this conversation (irrespective of how the conversation is mediated, e.g., by an interviewer or a computer). To minimize response errors, questionnaires should be crafted in accordance with best practices.

Recommendations about best practices stem from experience and common lore, on the one hand, and methodological research, on the other. In this chapter, we first offer recommendations about optimal questionnaire design based on the common wisdom (focusing mainly on the words used in questions), and then make further recommendations based on a review of the methodological research (focusing mainly on the structural features of questions).

We begin our examination of the methodological research by considering open versus closed questions, a difference especially relevant to three types of measurement: (1) asking for choices among nominal categories (e.g., "What is the most important problem facing the country?" (2) ascertaining numeric quantities (e.g., "How many hours did you watch television last week?") and (3) testing factual knowledge (e.g., "Who is Joseph Biden?").

Next, we treat the design of rating scales. We review the literature on the optimal number of scale points, consider whether some or all scale points should be labeled with words and/or numbers, and examine the problem of acquiescence response bias and methods for avoiding it. We then turn to the impact of response option order, outlining how it varies depending on whether categories are nominal or ordinal and whether they are presented visually or orally.

After that, we assess whether to offer "don't know" or no-opinion among a question's

explicit response options. Then we discuss social desirability response bias (a form of motivated misreporting) and recall bias (a form of unmotivated misreporting), and recommend ways to minimize each. Finally, we consider the ordering of questions and conclude with a discussion of how to test and evaluate questions via pretesting.

Conventional Wisdom

Hundreds of methodology textbooks have offered various versions of conventional wisdom about optimal question design. The most valuable advice in this common wisdom can be summarized as follows:

1. Use simple, familiar words (avoid technical terms, jargon, and slang);
2. Use simple syntax;
3. Avoid words with ambiguous meanings, i.e., aim for wording that all respondents will interpret in the same way;
4. Strive for wording that is specific and concrete (as opposed to general and abstract);
5. Make response options exhaustive and mutually exclusive;
6. Avoid leading or loaded questions that push respondents toward an answer;
7. Ask about one thing at a time (avoid double-barreled questions); and
8. Avoid questions with single or double negations.

Conventional wisdom also contains advice about how to optimize question order:

1. Early questions should be easy and pleasant to answer, and should build rapport between the respondent and the researcher;
2. Questions at the very beginning of a questionnaire should explicitly address the topic of the survey, as it was described to the respondent prior to the interview;
3. Questions on the same topic should be grouped together;

4. Questions on the same topic should proceed from general to specific;
5. Questions on sensitive topics that might make respondents uncomfortable should be placed at the end of the questionnaire; and
6. Filter questions should be included, to avoid asking respondents questions that do not apply to them.

Finally, the conventional wisdom recommends pretesting questionnaires, though it has little to say about how this is best accomplished.

Taken together these recommendations are of great value, but there is even more to be learned from the results of methodological research.

Optimizing vs. Satisficing

There is widespread agreement about the cognitive processes involved in answering questions optimally (e.g., Cannell, Miller, & Oksenberg 1981; Schwarz & Strack 1985; Tourangeau & Rasinski 1988). Specifically, respondents are presumed to execute each of four steps. First, they must interpret the question and deduce its intent. Next, they must search their memories for relevant information, and then integrate whatever information comes to mind into a single judgment. Finally, they must translate the judgment into a response, by selecting one of the alternatives offered by the question.

Each of these steps can be quite complex, involving considerable cognitive work (see Tourangeau and Bradburn, this volume). A wide variety of motives may encourage respondents to do this work, including desires for self-expression, interpersonal response, intellectual challenge, self-understanding, altruism, or emotional catharsis (see Warwick & Lininger 1975, pp. 185-187). Effort can also be motivated by the desire to assist the survey sponsor, e.g., to help employers improve working conditions, businesses design better products, or governments make

better-informed policy. To the extent that such motives inspire a respondent to perform the necessary cognitive tasks in a thorough and unbiased manner, the respondent may be said to be *optimizing*.

As much as we hope all respondents will optimize throughout a questionnaire, this is often an unrealistic expectation. Some people may agree to complete a questionnaire as result of a relatively automatic compliance process (see, e.g., Cialdini 1993) or because they are required to do so. Thus, they may agree merely to provide answers, with no intrinsic motivation to make the answers of high quality. Other respondents may satisfy whatever desires motivated them to participate after answering a first set of questions, and become fatigued, disinterested, or distracted as a questionnaire progresses further.

Rather than make the effort necessary to provide optimal answers, respondents may take subtle or dramatic shortcuts. In the former case, respondents may simply be less thorough in comprehension, retrieval, judgment, and response selection. They may be less thoughtful about a question's meaning; search their memories less comprehensively; integrate retrieved information less carefully; or select a response choice less precisely. All four steps are executed, but less diligently than when optimizing occurs. Instead of attempting the most accurate answers, respondents settle for merely satisfactory answers. The first answer a respondent considers that seems acceptable is the one offered. This response behavior might be termed weak *satisficing* (Krosnick 1991, borrowing the term from Simon 1957).

A more dramatic shortcut is to skip the retrieval and judgment steps altogether. That is, respondents may interpret each question superficially and select what they believe will appear to be a reasonable answer. The answer is selected without reference to any internal psychological cues specifically relevant to the attitude, belief, or event of interest. Instead, the respondent may

look to the wording of the question for a cue, pointing to a response that can be easily selected and easily defended if necessary. If no such cue is present, the respondent may select an answer completely arbitrarily. This process might be termed *strong satisficing*.

It is useful to think of optimizing and strong satisficing as the two ends of a continuum indicating the degrees of thoroughness with which the four response steps are performed. The optimizing end of the continuum involves complete and effortful execution of all four steps. The strong satisficing end involves little effort in the interpretation and answer reporting steps and no retrieval or integration at all. In between are intermediate levels.

The likelihood of satisficing is thought to be determined by three major factors: task difficulty, respondent ability, and respondent motivation (Krosnick 1991). Task difficulty is a function of both question-specific attributes (e.g., the difficulty of interpreting a question and of retrieving and manipulating the requested information) and attributes of the questionnaire's administration (e.g., the pace at which an interviewer reads the questions and the presence of distracting events). Ability is shaped by the extent to which respondents are adept at performing complex mental operations, practiced at thinking about the topic of a particular question, and equipped with pre-formulated judgments on the issue in question. Motivation is influenced by need for cognition (Cacioppo, Petty, Feinstein, & Jarvis 1996), the degree to which the topic of a question is personally important, beliefs about whether the survey will have useful consequences, respondent fatigue, and aspects of questionnaire administration (such as interviewer behavior) that either encourage optimizing or suggest that careful reporting is not necessary.

Efforts to minimize task difficulty and maximize respondent motivation are likely to pay off by minimizing satisficing and maximizing the accuracy of self-reports. As we shall see, the

notion of satisficing is useful for understanding why some questionnaire design decisions can improve the quality of answers.

Open versus Closed Questions

One of the first decisions a researcher must make when designing a survey question is whether to make it open (permitting respondents to answer in their own words) or closed (requiring respondents to select an answer from a set of choices). Although the vast majority of survey questions are closed, some open questions play prominent roles in survey research, for instance, those about the most important problem facing the country.

In order to analyze the answers to open questions, they must be grouped into a relatively small number of categories. This requires the development of a coding scheme; its application by more than one person; and the attainment of a high level of agreement between coders. The costs of these procedures, coupled with both the difficulties interviewers confront in recording open answers and the longer interview time taken by open questions, are responsible for the widespread use of closed questions.

These practical disadvantages of open questions, however, do not apply to the measurement of quantities. The answer categories to open questions about amounts -- for instance, number of doctor visits, hours devoted to housework, dollars spent for a good -- are implicit in the question, so no coding is required, and no special burden is placed on interviewers. Moreover, offering respondents a set of closed quantity categories (e.g. less than 1 hour, 1 to 3 hours, more than 3 hours) can produce error. Evidence indicates that the way in which amounts are divided to form closed categories conveys information that may bias respondent answers (Schwarz et al., 1985). Thus open questions are usually preferable to closed

items for measuring quantities.[1]

In measuring categorical judgments (such as the "most important problem"), where the options represent different objects, as opposed to points along a single continuum, researchers sometimes try to combine open and closed formats by including an "other" response alternative in addition to specifying a set of substantive choices. This is generally not effective, however, as respondents tend to restrict their answers to the substantive choices that are explicitly offered (Lindzey & Guest, 1951; Schuman and Scott, 1987).

If the list of choices offered by a closed question omits objects that a significant number of respondents would have mentioned to an open form of the question, even the rank ordering of the objects can differ across versions of the question. Therefore, a closed categorical question can often be used only if its answer choices are comprehensive. In some cases, identifying these categories will require a large-scale pretest of an open version of the question. In such instances, it may be more practical simply to ask an open question than to do the necessary pretesting.

Open and closed questions may also differ in their ability to measure possession of factually-correct knowledge. Closed questions will generally suffer more than open questions from correct guessing, though statistical adjustments to multi-item tests can adjust for this. Consistent with this logic, Krosnick and Fabrigar's (forthcoming) review of student testing studies indicates that open items provide more reliable and valid measurement than do closed items. On the other hand, open questions might be more likely than closed questions to elicit "don't know" (DK) answers from people who know the correct answer but are not sure they do

[1] Two reservations sometimes expressed about measuring quantities with open questions are that some respondents will say they don't know or refuse to answer and others will round their answers. In order to minimize missing data, respondents who do not give an amount to the open question can be asked follow-up closed questions, such as "Was it more or less than X?" (see, for example, Juster and Smith, 1997). Minimizing rounded answers is more difficult, but the problem may apply as much to closed questions as to open.

(and therefore decline to speculate in order to avoid embarrassment) or because they do not immediately recall the answer (and want to avoid expending the effort required to retrieve or infer it). In line with this speculation, Mondak (2001) found that open questions measuring political knowledge were more valid when DKs were discouraged than when they were encouraged in a nationwide survey of adults. Open questions may be more likely to elicit such illusory "don't know" responses in general population surveys than in tests administered to students in school (who would presumably be more motivated to guess or to work at generating an answer, since their grade hinges on it). So open knowledge questions may only perform well in surveys if DK responses are discouraged and guessing is encouraged. These issues merit more careful study with general population samples.

Open questions can add richness to survey results that is difficult, if not impossible, to achieve with closed questions, so including some (on their own or as follow-ups to closed items) can yield significant benefit (Schuman 1972).[2]

Number of Points on Rating Scales

When designing a rating scale, a researcher must specify the number of points on the scale. Likert (1932) scaling most often uses 5 points; Osgood, Suci, and Tannenbaum's (1957) semantic differential uses 7 points; and Thurstone's (1928) equal-appearing interval method uses 11 points. The American National Election Study surveys have measured citizens' political attitudes over the last 60 years using 2-, 3-, 4-, 5-, 7-, and 101-point scales (Miller, 1982). Robinson, Shaver, and Wrightsman's (1999) catalog of rating scales for a range of social

[2] Paradoxically, the openness of open questions can sometimes lead to narrower interpretations than comparable closed questions. Schuman and Presser (1981), for instance, found that an open version of the most important problem facing the nation question yielded many fewer "crime and violence" responses than a closed version that offered that option, presumably because respondents thought of crime as a local (as opposed to national) problem on the open version but not on the closed. The specificity resulting from the inclusion of response options can be an important advantage of closed questions.

psychological constructs and political attitudes describes 37 using 2-point scales, 7 using 3-point scales, 10 using 4-point scales, 27 using 5-point scales, 6 using 6-point scales, 21 using 7-point scales, two using 9-point scales, and one using a 10-point scale. Rating scales used to measure public approval of the U.S. President's job performance vary from 2 points to 5 points (Morin, 1993; Sussman, 1978). Thus, there appears to be no standard for the number of points on rating scales, and common practice varies widely.

In fact, however, the literature suggests that some scale lengths are preferable to maximize reliability and validity. In reviewing this literature, we begin with a discussion of theoretical issues and then describe the findings of relevant empirical studies.

Theoretical Issues

Respondents confronted with a rating scale must execute a matching or mapping process. They must assess their own attitude in conceptual terms (e.g., "I like it a lot") and then find the point on the rating scale that most closely matches that attitude (see Ostrom & Gannon, 1996). Thus, several conditions must be met in order for a rating scale to work effectively. First, the points offered should cover the entire measurement continuum, leaving out no regions. Second, these points must appear to be ordinal, progressing from one end of a continuum to the other, and the meanings of adjacent points should not overlap. Third, each respondent must have a relatively precise and stable understanding of the meaning of each point on the scale. Fourth, most or all respondents must agree in their interpretations of the meanings of each scale point. And a researcher must know what those interpretations are.

If some of these conditions are not met, data quality is likely to suffer. For example, if respondents fall in a particular region of an underlying evaluative dimension (e.g., "like somewhat") but no response options are offered in this region (e.g., a scale composed only of

"dislike" and "like"), they will be unable to rate themselves accurately. If respondents interpret the points on a scale one way today and differently next month, then they may respond differently at the two times, even if their underlying attitude has not changed. If two or more points on a scale appear to have the same meaning (e.g., "some of the time" and "occasionally") respondents may be puzzled about which one to select, leaving them open to making an arbitrary choice. If two people differ in their interpretations of the points on a scale, they may give different responses even though they may have identical underlying attitudes. And if respondents interpret scale point meanings differently than researchers do, the researchers may assign numbers to the scale points for statistical analysis that misrepresent the messages respondents attempted to send via their ratings.

Translation ease. The length of scales can impact the process by which people map their attitudes onto the response alternatives. The ease of this mapping or translation process varies, partly depending upon the judgment being reported. For instance, if an individual has an extremely positive or negative attitude toward an object, a dichotomous scale (e.g., "like," "dislike") easily permits reporting that attitude. But for someone with a neutral attitude, a dichotomous scale without a midpoint would be suboptimal, because it does not offer the point most obviously needed to permit accurate mapping.

A trichotomous scale (e.g., "like," "neutral," "dislike") may be problematic for another person who has a moderately positive or negative attitude, equally far from the midpoint and the extreme end of the underlying continuum. Adding a moderate point on the negative side (e.g., "dislike somewhat") and one on the positive side of the scale (e.g., "like somewhat") would solve this problem. Thus, individuals who want to report neutral, moderate, or extreme attitudes would all have opportunities for accurate mapping.

The value of adding even more points to a rating scale may depend upon how refined people's mental representations of the construct are. Although a 5-point scale might be adequate, people may routinely make more fine-grained distinctions. For example, most people may be able to differentiate feeling slightly favorable, moderately favorable, and extremely favorable toward objects, in which case a 7-point scale would be more desirable than a 5-point scale.

If people do make fine distinctions, potential information gain increases as the number of scale points increases, because of greater differentiation in the judgments made (for a review, see Alwin, 1992). This will be true, however, only if individuals do in fact make use of the full scale, which may not occur with long scales.

The ease of mapping a judgment onto a response scale is likely to be determined in part by how close the judgment is to the conceptual divisions between adjacent points on the scale. For example, when people with an extremely negative attitude are asked, "Is your opinion of the President very negative, slightly negative, neutral, slightly positive, or very positive?" they can easily answer "very negative," because their attitude is far from the conceptual division between "very negative" and "slightly negative." However, individuals who are moderately negative have a true attitude close to the conceptual division between "very negative" and "slightly negative," so they may face a greater challenge in using this 5-point rating scale. The "nearness" of someone's true judgment to the nearest conceptual division between adjacent scale points is associated with unreliability of responses – those nearer to a division are more likely to pick one option on one occasion and another option on a different occasion (Kuncel, 1973, 1977).

Clarity of scale point meanings. In order for ratings to be reliable, people must have a clear understanding of the meanings of the points on the scale. If the meaning of scale points is ambiguous, then both reliability and validity of measurement may be compromised.

A priori, it seems that dichotomous response option pairs are very clear in meaning; that is, there is likely to be considerable consensus on the meaning of options such as "favor" and "oppose" or "agree" and "disagree." Clarity may be compromised when a dichotomous scale becomes longer, because each point added is one more point to be interpreted. And the more such interpretations a person must make, the more chance there is for inconsistency over time or across individuals. That is, it is presumably easier for someone to identify the conceptual divisions between "favoring," "opposing," and being "neutral" on a trichotomous item than on a seven-point scale, where six conceptual divisions must be specified.

For rating scales up to seven points long, it may be easy to specify intended meanings of points with words, such as "like a great deal," "like a moderate amount," "like a little," "neither like nor dislike," "dislike a little," "dislike a moderate amount," and "dislike a great deal." But once the number of scale points increases above seven, point meanings may become considerably less clear. For example, on 101-point attitude scales (sometimes called feeling thermometers), what exactly do 76, 77, and 78 mean? Even for 11- or 13-point scales, people may be hard-pressed to define the meaning of the scale points.

Uniformity of scale point meaning. The number of scale points used is inherently confounded with the extent of verbal labeling possible, and this confounding may affect uniformity of interpretations of scale point meanings across people. Every dichotomous and trichotomous scale must, of necessity, include verbal labels on all scale points, thus enhancing their clarity. But when scales have four or more points, it is possible to label only the end points with words. In such cases, comparisons with dichotomous or trichotomous scales reflect the impact of both number of scale points and verbal labeling. It is possible to provide an effective verbal label for each point on a scale containing more than 7 points, but doing so becomes more

difficult as the number of scale points increases beyond that length.

The respondent's task may be made more difficult when presented with numerical rather than verbal labels. To make sense of a numerically-labeled rating scale, respondents must first generate a verbal definition for each point and then match these definitions against their mental representation of the attitude of interest. Verbal labels might therefore be advantageous, because they may clarify the meanings of the scale points while at the same time reducing respondent burden by removing a step from the cognitive processes entailed in answering the question.

Satisficing. Finally, the optimal number of rating scale points may depend on individuals' cognitive skills and motivation to provide accurate reports. Offering a midpoint on a scale may constitute a cue encouraging satisficing to people low in ability and/or motivation, especially if its meaning is clearly either "neutral/no preference" or "status quo - keep things as they are now." If pressed to explain these answers, satisficing respondents might have little difficulty defending such replies. Consequently, offering a mid-point may encourage satisficing by providing a clear cue offering an avenue for doing so.

However, there is a potential cost to eliminating midpoints. Some people may truly belong at the scale midpoint and may wish to select such an option to communicate their genuine neutrality or endorsement of the status quo. If many people have neutral attitudes to report, eliminating the midpoint will force them to pick a point either on the positive side or on the negative side of the scale, resulting in inaccurate measurement.

The number of points on a rating scale can also impact satisficing via a different route: task difficulty. The number of scale points offered on a rating scale may be a determinant of task difficulty. Two-point scales simply require a decision of direction (e.g., pro versus con), whereas longer scales require decisions of both direction and extremity. Very long scales require people

to choose between many options, so these scales may be especially difficult in terms of scale point interpretation and mapping. Yet providing too few scale points may contribute to task difficulty by making it impossible to express moderate positions. Consequently, task difficulty (and satisficing as well) may be at a minimum for moderately long rating scales, resulting in more accurate responses.

Evidence on the Optimal Number of Scale Points

Many investigations have produced evidence useful for inferring the optimal number of points on rating scales. Some of this work has systematically varied the number of scale points offered while holding constant all other aspects of questions. Other work has attempted to discern people's natural discrimination tendencies in using rating scales. Several of the studies we review did not explicitly set out to compare reliability or validity of measurement across scale lengths but instead reported data that permit us to make such comparisons post hoc.

Reliability. Lissitz and Green (1975) explored the relation of number of scale points to reliability using simulations. These investigators generated sets of true attitudes and random errors for groups of hypothetical respondents and then added these components to generate responses to attitude questions on different-length scales in two hypothetical "waves" of data. Cross-sectional and test-retest reliability increased from 2- to 3- to 5-point scales but were equivalent thereafter for 7-, 9-, and 14-point scales. Similar results were obtained in simulations by Jenkins and Taber (1977), Martin (1978), and Srinivasan and Basu (1989).

Some studies have found the number of scale points to be unrelated to cross-sectional reliability. Bendig (1954) found that ratings using either 2-, 3-, 5-, 7-, or 9-point scales were equivalently reliable. Similar results have been reported for scales ranging from 2 to 7 points (Komorita & Graham, 1965; Masters, 1974) and for longer scales ranging from 2 to 19 points

(Birkett, 1986; Matell & Jacoby, 1971; Jacoby & Matell, 1971). Other studies have yielded differences that are consistent with the notion that scales of intermediate lengths are optimal (Birkett, 1986; Givon & Shapira, 1984; Masters, 1974). For example, Givon and Shapira (1984) found pronounced improvements in item reliability when moving from 2-point scales toward 7-point scales. Reliability continued to increase up to lengths of 11 points, but the increases beyond 7 points were quite minimal for single items.

Another way to assess optimal scale length is to collect data on a scale with many points and recode it into a scale with fewer points. If longer scales contain more random measurement error, then recoding should improve reliability. But if longer scales contain valid information that is lost in the recoding process, then recoding should reduce data quality. Consistent with this latter hypothesis, Komorita (1963) found that cross-sectional reliability for 6-point scales was 0.83, but only 0.71 when the items were recoded to be dichotomous. Thus, it appears that more reliable information was contained in the full 6-point ratings than the dichotomies. Similar findings were reported by Matell and Jacoby (1971) indicating that collapsing scales longer than 3 points discarded reliable information, because long scales provided more information than short scales and were no less reliable.

Although there is some variation in the patterns yielded by these studies, they generally support the notion that reliability is lower for scales with only two or three points compared to those with more points, but suggest that the gain in reliability levels off after about 7 points.

Validity. Studies estimating correlations between true attitude scores and observed ratings on scales of different lengths using simulated data have found that validity increases as scales lengthen from 2 points; however as scales grow longer, the gains in validity become correspondingly smaller (Green & Rao, 1970; Lehmann & Hulbert, 1972; Lissitz & Green, 1975;

Martin, 1973; Martin, 1978; Ramsay, 1973).

Other techniques to assess the validity of scales of different lengths have included: correlating responses obtained from two different ratings of the same construct (e.g., Matell & Jacoby, 1971; Smith, 1994a; Smith & Peterson, 1985; Watson, 1988; Warr, Barter, & Brownridge, 1983), correlating attitude measures obtained using scales of different lengths with other attitudes (e.g., Schuman & Presser, 1981: 175-176), and using the ratings obtained using different scale lengths to predict other attitudes (Rosenstone, Hansen, & Kinder, 1986; Smith & Peterson, 1985). These studies have typically found that concurrent validity improves with increasing scale length.

Several studies suggest that longer scales are less susceptible to question order effects (Wedell & Parducci, 1988; Wedell, Parducci, & Lane, 1990; Wedell, Parducci, & Geiselman, 1987). However, one study indicates that especially long scales might be more susceptible to context effects than those of moderate length (Schwarz & Wyer, 1985). Stember and Hyman (1949/1950) found that answers to dichotomous questions were influenced by interviewer opinion, but this influence disappeared among individuals who were also offered a middle alternative, yielding a trichotomous question.

As with the research on reliability, these studies generally support the notion that validity is higher for scales with a moderate number of points than for scales with fewer, with the suggestion that validity is compromised by especially long scales.

Discerning natural scale differentiation. In a study by Champney and Marshall (1939), judges provided ratings on various scales by placing "x"s on 9-centimeter-long lines. Five, six, or seven points along the lines were labeled with sentences to establish the meanings of the parts of the scale. The continuous measurement procedure allowed Champney and Marshall (1939) to

divide the lines into as many equally-sized categories as they wished and then assess the cross-sectional reliability of the various divisions for two items that were both designed to measure sociability. Cross-sectional reliability increased dramatically from a 2-point scale ($r = 0.56$) to a 9-point scale ($r = 0.70$), and a further significant increase appeared when moving to 18 scale points ($r = 0.74$). Reliabilities, however, were essentially the same for 22 ($r = 0.75$), 30 ($r = 0.76$), 45 points ($r = 0.77$), and 90 points ($r = 0.76$). The judges returned three weeks later to re-rate the objects on a total of 12 scales, which allowed the computation of test-retest reliability of ratings, and results were consistent with the cross-sectional findings.

McKelvie (1978) had subjects rate various objects by marking points on lines with no discrete category divisions. The subjects also indicated their "confidence interval" around each judgment. By dividing the total line length by the average magnitude of the confidence interval, McKelvie (1978) could estimate the number of scale points subjects were naturally employing, which turned out to be 5.

Another study along these lines examined the number of scale points that people used on scales of increasing length. Matell and Jacoby (1972) had individuals provide a series of ratings on scales of lengths ranging from 2 points to 19 points. Nearly everyone used both points on the dichotomous items, and most people used all three points on the trichotomous items. For longer scales, people used about half the points offered, regardless of length. That is, the more scale points that were offered up to 19, the more points people used, up to about 9.

Rundquist and Sletto (1936) had subjects complete a set of ratings either by marking points on lines or by using 5- or 7-point category scales. When the line marks were coded according to a 7-point division, the distribution of ratings was identical to that from the 7-point scale. But when the line marks were coded according to a 5-point division, the distribution was

significantly different from the 5-point scale, with fewer extreme and midpoint ratings being made for the latter than the former.

Middle alternatives and satisficing. The relevance of the satisficing perspective to middle alternatives can be gauged by determining whether respondents are most attracted to them under the conditions that are thought to foster satisficing, two of which are low cognitive skills and low attitude strength (see Krosnick, 1991). Kalton, Roberts, and Holt (1980), Schuman and Presser (1981), O'Muircheartaigh, Krosnick, and Helic (1999), and Narayan and Krosnick (1996) concluded that attraction to middle alternatives was unrelated to educational attainment (a proxy measure for cognitive skills). Krosnick and Schuman (1988) and Bishop (1990) found more attraction among those for whom the issue was less important and whose attitudes were less intense, and O'Muircheartaigh et al. (1999) found that attraction to middle alternatives was greater among people with less interest in the topic. But Stember and Hyman (1949/1950) found attraction to middle alternatives on a specific policy issue was unrelated to general interest in foreign policy, and O'Muircheartaigh et al. (1999) found no relation of attraction to middle alternatives with volume of knowledge about the object. Thus, the evidence on the connection between middle alternatives and satisficing is mixed.

More importantly, O'Muircheartaigh and colleagues (1999) found that adding midpoints to rating scales improved the reliability and validity of ratings. Structural equation modeling of error structures revealed that omitting the middle alternative led respondents to randomly select one of the moderate scale points closest to where a midpoint would appear. This suggests that offering midpoints is desirable.[3]

[3] Almost all the studies reviewed above involved experimental designs varying the number of rating scale points, holding constant all other aspects of the questions. Some additional studies have explored the impact of number of scale points using a different approach: Meta-analysis. These studies have taken large sets of questions asked in pre-existing surveys, estimated their reliability and/or validity, and meta-analyzed the results to see whether

Overall, our review suggests that 7-point scales are probably optimal in many instances. However, it is important to note that most of the literature on number of points involves visual administration. Thus there is some uncertainty about its applicability to telephone surveys. This is especially so given that oral presentation of 7-point scales on the telephone may require branching, i.e., the conversion of one question into two. Nonetheless, Krosnick and Berent (1993) found that a two-item branching format took less time in a telephone survey than the equivalent one-item 7 point scale.

Labeling of Rating Scale Points

Once the length of a rating scale has been specified, a researcher must decide how to label the points. Various studies suggest that reliability is higher when all points are labeled with words than when only some are (e.g., Krosnick & Berent, 1993). Respondents also express greater satisfaction when more scale points are verbally labeled (e.g., Dickinson & Zellinger, 1980). Researchers can maximize reliability and validity by selecting labels that divide up the continuum into approximately equal units (e.g., Klockars & Yamagishi, 1988; for a summary, see Krosnick & Fabrigar, forthcoming).[4]

Many closed attitude measures are modeled after Likert's technique, offering statements to respondents and asking them to indicate whether they agree or disagree with each or to

---

data quality varies with scale point number (e.g., Alwin, 1992, 1997; Alwin & Krosnick, 1991; Andrews, 1984, 1990; Scherpenzeel, 1995). However, these meta-analyses sometimes mixed together measures of subjective judgments with measurements of objective constructs such as numeric behavior frequencies (e.g., number of days) and routinely involved strong confounds between number of scale points and other item characteristics, only some of which were measured and controlled for statistically. Consequently, it is not surprising that these studies yielded inconsistent findings. For example, Andrews (1984) found that validity and reliability were worst for 3-point scales, better for 2-point and 4-point scales, and even better as scale length increased from 5 points to 19 points. In contrast, Alwin and Krosnick (1991) found that 3-point scales had the lowest reliability, found no difference in the reliabilities of 2-, 4-, 5, and 7-point scales, and found 9-point scales to have maximum reliability (though these latter scales actually offered 101 response alternatives). And Scherpenzeel (1995) found the highest reliability for 4/5-point scales, lower reliability for 10 points, and even lower for 100 points. We therefore view these studies as less informative than experiments that manipulate rating scale length.

[4] This suggests that analog devices such as thermometers or ladders may not be good measuring devices.

indicate their level of agreement or disagreement. Other attitude measures offer assertions and ask people to report the extent to which the assertions are true or false, and some attitude measures ask people "yes/no" questions (e.g., "Do you favor limiting imports of foreign steel?").

These sorts of item formats are 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 perceived goodness, the same set of response options can be used.

Nonetheless, these question formats may be problematic. People may sometimes say "agree," "true," or "yes" regardless of the question being asked of them. For example, a respondent might agree with the statement that "individuals are mainly to blame for crime" and also agree with the statement that "social conditions are mainly to blame for crime." This behavior, labeled "acquiescence," can be defined as endorsement of an assertion made in a question, regardless of the assertion's content. The behavior 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).

The evidence documenting acquiescence by a range of methods is now voluminous, (for a review, see Krosnick & Fabrigar, forthcoming). Consider first agree/disagree questions. When people are given the choices "agree" and "disagree," are not told the statements to which they

apply, and are asked to guess what answers an experimenter is imagining, "agree" is chosen much more often than "disagree" (e.g., Berg & Rapaport, 1954). When people are asked to agree or disagree with pairs of statements stating mutually exclusive views (e.g., "I enjoy socializing" vs. "I don't enjoy socializing"), the between-pair correlations are negative but generally very weakly so (Krosnick and Fabrigar report an average correlation of only -0.22 across 41 studies). Although random measurement error could cause the correlations to depart substantially from –1.0, acquiescence could do so as well.

Consistent with this possibility, averaging across 10 studies, 52% of people agreed with an assertion, whereas only 42% of people disagreed with the opposite assertion (Krosnick & Fabrigar, forthcoming). 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. On average 14% more people agreed with an assertion than expressed the same view in the corresponding forced choice question. In 7 other studies, an average of 22% of the respondents agreed with both a statement and its reversal, whereas only 10% disagreed with both. Thus, taken together, these methods suggest an acquiescence effect averaging about 10%.

Other evidence indicates that the tendency to acquiesce is a general inclination of some individuals across questions. The cross-sectional reliability of the tendency to agree with assertions averaged 0.65 across 29 studies. And the over-time consistency of the tendency to acquiesce was about 0.75 over one month, 0.67 over four months, and 0.35 over 4 years (e.g., Couch & Keniston, 1960; Hoffman, 1960; Newcomb, 1943).

Similar results (regarding correlations between opposite assertions, endorsement rates of items, their reversals, and 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 responses to these items as well (see Krosnick and Fabrigar, forthcoming). And there is other such evidence regarding these response alternatives. 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."

Acquiescence is most common among respondents who have lower social status (e.g., Gove & Geerken, 1977; Lenski & Leggett, 1960), less formal education (e.g., Ayidiya & McClendon, 1990; Narayan & Krosnick, 1996), lower intelligence (e.g., Forehand, 1962; Hanley, 1959; Krosnick, Narayan, & Smith, 1996), lower cognitive energy (Jackson, 1959), less enjoyment from thinking (Messick & Frederiksen, 1958), and less concern to convey a socially desirable image of themselves (e.g., Goldsmith, 1987; Shaffer, 1963). Also, acquiescence is most common when a question is difficult (Gage et al., 1957; Hanley, 1962; Trott & Jackson, 1967), when respondents have become fatigued by answering many prior questions (e.g., Clancy & Wachsler, 1971), and when interviews are conducted by telephone as opposed to face-to-face (e.g., Calsyn, Roader, & Calsyn, 1992 Holbrook, Green, & Krosnick, 2003). 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.

If this interpretation is correct, acquiescence might be reduced by assuring (through pretesting) that questions are easy for people to comprehend and answer and by taking steps to maximize respondent motivation to answer carefully and thoughtfully. However, no evidence is yet available on whether acquiescence can be reduced in these ways. Therefore, a better approach to eliminating acquiescence is to avoid using agree/disagree, true/false, and yes/no

questions altogether. This is especially sensible because answers to these sorts of questions are less valid and less reliable than answers to the "same" questions expressed in a format that offers competing points of view and asks people to choose among them (e.g., Eurich, 1931; Isard, 1956; Watson & Crawford, 1930).

One alternative approach to controlling for acquiescence is derived from the presumption that certain people have acquiescent personalities and are likely to do all of the acquiescing. According to this view, a researcher 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 on moderating factors (e.g., position in the questionnaire and mode of administration) that we reviewed above suggests that acquiescence is not simply the result of having an acquiescent personality; rather, it is influenced by circumstantial factors. Because this "correction" approach does not take that into account, the corrections performed are not likely to fully adjust for acquiescence.

It might seem that acquiescence can be controlled by 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 Paulhus, 1991). This approach is designed to place acquiescers in the middle of the dimension but it will do so only if the assertions made in the reversals are as extreme as the original statements. 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 if one is able

to construct appropriately reversed items, acquiescers presumably end up at a point on the measurement dimension where most probably do not belong on substantive grounds. That is, if these individuals were induced not to acquiesce but to answer the items thoughtfully, their final scores would presumably be more valid than placing them at or near the midpoint of the dimension.

Most important, answering an agree/disagree, true/false, or yes/no question always requires respondents to first answer a comparable rating question with construct-specific response options. For example, people asked to agree or disagree with the assertion "I like bananas," must first decide how positive or negative their attitudes are toward bananas (perhaps concluding "I love bananas") and then translate that conclusion into the appropriate selection in order to answer the question. Researchers who use such questions presume that arraying people along the agree/disagree dimension corresponds monotonically to arraying them along the underlying substantive dimension of interest. That is, the more people agree with the assertion "I like bananas," the more positive is their true attitude toward bananas.

Yet consider respondents asked for their agreement with the statement "I am usually pretty calm." They may "disagree" because they believe they are always very calm or because they are never calm, which violates the monotonic equivalence of the response dimension and the underlying construct of interest. As this example makes clear, it would be simpler to ask people directly about the underlying dimension. Every agree/disagree, true/false, or yes/no question implicitly requires the respondent to rate an object along a continuous dimension, so asking about that dimension directly is bound to be less burdensome. Not surprisingly, then, the reliability and validity of rating scale questions that array the full attitude dimension explicitly (e.g., from "extremely bad" to "extremely good," or from "dislike a great deal" to "like a great

deal") are higher than those of agree/disagree, true/false, and yes/no questions that focus on only a single point of view (e.g., Ebel, 1982; Mirowsky & Ross, 1991; Ruch & DeGraff, 1926; Saris & Krosnick, 2000; Wesman, 1946). Consequently, it seems best to avoid agree/disagree, true/false, and yes/no formats altogether and instead ask questions using rating scales that explicitly display the evaluative dimension.

The Order of Response Alternatives

Many studies have shown that the order in which response alternatives are presented can affect their selection. Some studies show primacy effects (options more likely to be selected when they are presented early); others show recency effects (options more likely to be selected when presented last), and still other studies show no order effects at all. Satisficing theory helps explain the apparent contradictions.

We consider first how response order affects categorical questions and then turn to its effect in rating scales. Response order effects in categorical questions (e.g., "Which do you like more, peas or carrots?") appear to be attributable to "weak satisficing." When confronted with categorical questions, optimal answering would entail carefully assessing the appropriateness of each of the offered response alternatives before selecting one. In contrast, a weak satisficer would simply choose the first response alternative that appears to constitute a reasonable answer. Exactly which alternative is most likely to be chosen depends in part upon whether the choices are presented visually or orally.

When categorical alternatives are presented visually, either on a show-card in a face-to-face interview or in a self-administered questionnaire, weak satisficing is likely to bias respondents toward selecting choices displayed early in a list. Respondents are apt to consider each alternative individually beginning at the top of the list, and their thoughts are likely to be

biased in a confirmatory direction (Koriat, Lichtenstein, & Fischhoff 1980; Klayman & Ha 1984; Yzerbyt & Leyens 1991). Given that researchers typically include choices that are plausible, confirmation-biased thinking will often generate at least a reason or two in favor of most of the alternatives in a question.

After considering one or two alternatives, the potential for fatigue (and therefore reduced processing of later alternatives) is significant. Fatigue may also result from proactive interference, whereby thoughts about the initial alternatives interfere with thinking about later, competing alternatives (Miller & Campbell 1959). Weak satisficers cope by thinking only superficially about later alternatives; the confirmatory bias thereby advantages the earlier items. Alternatively, weak satisficers can simply terminate their evaluation altogether once they come upon an alternative that seems to be a reasonable answer. Because many answers are likely to seem reasonable, such respondents are again apt to end up choosing alternatives near the beginning of a list. Thus, weak satisficing seems liable to produce primacy effects under conditions of visual presentation.

When response alternatives are presented orally, as in face-to-face or telephone interviews, the effects of weak satisficing are more difficult to anticipate. This is so because order effects reflect not only evaluations of each option, but also the limits of memory. When categorical alternatives are read aloud, presentation of the second alternative terminates processing of the first one, usually relatively quickly. Therefore, respondents are able to devote the most processing time to the final items; these items remain in short-term memory after interviewers pause to let respondents answer.

It is conceivable that some people listen to a short list of categorical alternatives without evaluating any of them. Once the list is completed, these individuals may recall the first

alternative, think about it, and then progress forward through the list from there. Given that fatigue should instigate weak satisficing relatively quickly, a primacy effect would be expected. However, because this approach requires more effort than first considering the final items in the list, weak satisficers are unlikely to use it very often. Therefore, considering only the allocation of processing, we would anticipate both primacy and recency effects, though the latter should be more common than the former.

These effects of deeper processing are likely to be reinforced by the effects of memory. Categorical alternatives presented early in a list are most likely to enter long-term memory (e.g., Atkinson & Shiffrin 1968), and those presented at the end of a list are most likely to be in short-term memory immediately after the list is heard (e.g., Atkinson & Shiffrin 1968). Furthermore, options presented late are disproportionately likely to be recalled (Baddeley & Hitch 1977). So options presented at the beginning and end of a list are more likely to be recalled after the question is read, particularly if the list is long. Therefore, both early and late categorical options should be more available for selection, especially among weak satisficers. Short-term memory usually dominates long-term memory immediately after acquiring a list of information (Baddeley & Hitch 1977), so memory factors should promote recency effects more than primacy effects. Thus, in response to orally presented questions, mostly recency effects would be expected, though some primacy effects might occur as well.

Schwarz and Hippler (1991; Schwarz, Hippler, & Noelle-Neumann 1992) note two additional factors that may govern response order effects: the plausibility of the response alternatives presented, and perceptual contrast effects. If deep processing is accorded to an alternative that seems highly implausible, even people with a confirmatory bias in reasoning may

fail to generate any reasons to select it. Thus, deeper processing of some alternatives may make them especially unlikely to be selected.

Although studies of response order effects in categorical questions seem to offer a confusing pattern of results when considered as a group, a clearer pattern appears when the studies are separated into those involving visual and oral presentation. In visual presentation, primacy effects have been found (Ayidiya & McClendon 1990; Becker 1954; Bishop et al 1988; Campbell & Mohr 1950; Isreal & Taylor 1990; Krosnick & Alwin 1987; Schwarz, Hippler, & Noelle-Neumann 1992). In studies involving oral presentation, nearly all response order effects have been recency effects (McClendon 1986; Berg & Rapaport 1954; Bishop 1987; Bishop et al 1988; Cronbach 1950; Krosnick 1992; Krosnick & Schuman 1988; Mathews 1927; McClendon 1991; Rubin 1940; Schuman & Presser 1981; Schwarz, Hippler, & Noelle-Neumann 1992; Visser, Krosnick, Marquette, & Curtin, 1999).[5]

If the response order effects demonstrated in these studies are due to weak satisficing, then these effects should be stronger under conditions where satisficing is most likely. And indeed, these effects were stronger among respondents with relatively limited cognitive skills (Krosnick 1990; Krosnick & Alwin 1987; Krosnick, Narayan, & Smith 1996; McClendon 1986; McClendon 1991; Narayan & Krosnick 1996). Mathews (1927) also found stronger primacy effects as questions became more and more difficult and as people became more fatigued. And although McClendon (1986) found no relation between the number of words in a question and the magnitude of response order effects, Payne (1949/1950) found more response order effects in questions involving more words and words that were more difficult to comprehend. Also, Schwarz et al. (1992) showed that a strong recency effect was eliminated when prior questions

[5] Some studies have found no effect of response order. It is unclear what distinguishes them from the studies that do produce such effects.

on the same topic were asked, which presumably made knowledge of the topic more accessible and thereby made optimizing easier.

Much of the logic articulated above regarding categorical questions seems applicable to rating scales, but in a different way than for categorical questions. Many people's attitudes are probably not perceived as precise points on an underlying evaluative dimension but rather are seen as ranges or "latitudes of acceptance" (M. Sherif & Hovland 1961; C. W. Sherif, Sherif, & Nebergall 1965). If satisficing respondents consider the options on a rating scale sequentially, they may select the first one that falls in their latitude of acceptance, yielding a primacy effect under both visual and oral presentation.

Nearly all of the studies of response order effects in rating scales involved visual presentation, and when order effects appeared, they were almost uniformly primacy effects (Carp 1974; Chan 1991; Holmes 1974; Johnson 1981; Payne 1971; Quinn & Belson 1969). Furthermore, the two studies of rating scales that used oral presentation found primacy effects as well (Kalton et al 1978; Mingay & Greenwell 1989). Consistent with the satisficing notion, Mingay and Greenwell (1989) found that their primacy effect was stronger for people with more limited cognitive skills. However, these investigators found no relation of the magnitude of the primacy effect to the speed at which interviewers read questions, despite the fact that a fast pace presumably increased task difficulty. Also, response order effects were no stronger when questions were placed later in a questionnaire (Carp 1974). Thus, the moderators of rating scale response order effects may be different from the moderators of such effects in categorical questions, though more research is clearly needed to fully address this question.

How should researchers handle response order effects when designing survey questions? One seemingly effective way to do so is to counterbalance the order in which choices are

presented. Counterbalancing is relatively simple to accomplish with dichotomous questions; a random half of the respondents can be given one order, and the other half can be given the reverse order. When the number of response choices increases, the counterbalancing task can become more complex. However, when it comes to rating scales, it makes no sense to completely randomize the order in which scale points are presented, because that would eliminate the sensible progressive ordering from positive to negative, negative to positive, most to least, least to most, etc. Therefore, for scales, only two orders ought to be used, regardless of how many points are on the scale.

Unfortunately, counterbalancing order creates a new problem: variance in responses due to systematic measurement error. Once response alternative orders have been varied, respondent answers may differ from one another partly because different people received different orders. One might view this new variance as random error variance, the effect of which would be to attenuate observed relations among variables and leave marginal distributions of variables unaltered. However, given the theoretical explanations for response order effects, this error seems unlikely to be random.

Thus in addition to counterbalancing presentation order, it seems potentially valuable to take steps to reduce the likelihood of the effects occurring in the first place. The most effective method for doing so presumably depends on the cognitive mechanism producing the effect. If primacy effects are due to satisficing, then steps that reduce satisficing should reduce the effects. For example, with regard to motivation, questionnaires can be kept short, and accountability can be induced by occasionally asking respondents to justify their answers. With regard to task difficulty, the wording of questions and answer choices can be made as simple as possible.

Treatment of No-Opinion

What happens when people are asked a question about which they have no relevant knowledge? Ideally, they will say that they do not know the answer. But respondents may wish not to appear uninformed and may therefore give an answer to satisfy the interviewer (Converse, 1964). In order to reduce the likelihood of such behavior, some researchers have recommended that don't know (DK) options (or filters) routinely be included in questions (e.g., Bogart, 1972; Converse & Presser, 1986; Payne, 1950; Vaillancourt, 1973). By explicitly offering a DK option, filters tell respondents that it is acceptable to say they have no information with which to answer a question.

Do DK filters work? On the one hand, there is evidence that they successfully encourage people without information to admit it (Schuman and Presser, 1981). On the other hand, filters may go too far and discourage people who do have information with which to generate a meaningful answer from expressing it. In fact, there is considerable evidence that DK filters do not improve measurement.

Support for this conclusion comes from research that explored the meaningfulness of the substantive responses provided by people who would have said "don't know" if that option had been offered. Gilljam and Granberg (1993) asked three questions tapping attitudes toward building nuclear power plants. The first of these questions offered a DK option, and 15% selected it. The other two questions, asked later in the interview, did not offer DK options, and only 3% and 4%, respectively, failed to offer substantive responses to them. Thus, the majority of people who said "don't know" to the initial question offered opinions on the later two questions. Their later responses mostly reflected meaningful opinions, because they correlated moderately with one another and predicted the respondents' vote on a nuclear power referendum that occurred a few months after the interview.

Although Bishop, Oldendick, Tuchfarber, and Bennett (1979) found slightly stronger associations of attitudes with other criterion items when DK options were offered than when they were not, Schuman and Presser (1981) rarely found such differences. Alwin and Krosnick (1991), McClendon and Alwin (1993), Krosnick and Berent (1993), Krosnick et al. (2002), and Poe et al. (1988) found answers were no more reliable when DK filters were included in questions than when they were not.

Krosnick et al. (2002) also found that offering DK options did not enhance the degree to which people's answers were responsive to question manipulations that should have affected them. Respondents were asked whether they would be willing to pay a specified amount in additional taxes for an environmental program, with random subsamples offered or not offered an explicit DK. Random subsamples were also told different amounts, on the presumption that fewer people would be willing to pay for the program as the price increased. If failing to offer a DK option creates meaningless answers, then there would have been less price sensitivity among people pressed to offer substantive opinions than among those offered a DK option. But in fact, sensitivity to price was the same in both groups. Even more notably, Visser, Krosnick, Marquette, and Curtin (2000) found that pre-election polls predicted election outcomes more accurately when respondents who initially said they did not know were pressed to identify the candidate toward whom they leaned.

In interpreting these results on the effects of DK filters, it is useful to consider cognitive psychologists' work on the process by which people decide that they do not know something. Norman (1973) proposed a two-step model. If asked a question such as "Do you favor or oppose U.S. government aid to Nicaragua?" a respondent's first step would be to search long-term memory for any information relevant to the objects mentioned: U.S. foreign aid and Nicaragua.

If no information about either is recalled, the individual can quickly respond by saying “don’t know.” But if some information is located about either object, the person must then retrieve that information and decide whether it can be used to formulate a reasonable opinion. If not, the individual can then answer “don't know,” but the required search time makes this a relatively slow response. Glucksberg and McCloskey (1981) reported a series of studies demonstrating that “don't know” responses do indeed occur either quickly or slowly, the difference resulting from whether or not any relevant information can be retrieved in memory.

According to the proponents of DK filters, the most common reason for DKs is that the respondent lacks the necessary information and/or experience with which to form an attitude. This would presumably yield quick, first-stage DK responses. In contrast, second-stage DK responses could occur for other reasons, such as ambivalence: some respondents may know a great deal about an object and/or have strong feelings toward it, but their thoughts and/or feelings may be contradictory, making it difficult to select a single response.

DK responses might also result at the point at which respondents attempt to translate their judgment into the choices offered by a question. Thus people may know approximately where they fall on an attitude scale (e.g., around 6 or 7 on a 1-7 scale), but because of ambiguity in the meaning of the scale points or of their internal attitudinal cues, they may be unsure of exactly which point to choose, and therefore offer a DK response. Similarly, individuals who have some information about an object, have a neutral overall orientation toward it, and are asked a question without a neutral response option might say DK because the answer they would like to give has not been conferred legitimacy. Or people may be concerned that they do not know enough about the object to defend an opinion, so their opinion may be withheld rather than reported.

Finally, it seems possible that some DK responses occur before respondents have even

begun to attempt to retrieve relevant information. Thus respondents may say "don't know" because they do not understand the question (see, e.g., Fonda, 1951).

There is evidence that DK responses occur for all these reasons, but when people are asked directly why they say "don't know," they rarely mention lacking information or an opinion. Instead they most often cite the other reasons such as ambivalence (Coombs & Coombs, 1976; Faulkenberry & Mason, 1978; Klopfer & Madden, 1980; Schaeffer & Bradburn, 1989).

Satisficing theory also helps account for the fact that DK filters do not consistently improve data quality (Krosnick, 1991). According to this perspective, people have many latent attitudes that they are not immediately aware of holding. Because the bases of those opinions reside in memory, people can retrieve those bases and integrate them to yield an overall attitude, but doing so requires significant cognitive effort ("optimizing"). When people are disposed not to do this work and instead prefer to shortcut the effort of generating answers, they may attempt to satisfice by looking for question cues pointing to an acceptable answer that requires little effort to select. A DK option constitutes just such a cue and may therefore encourage satisficing, whereas omitting the DK option is more apt to encourage respondents to do the work necessary to retrieve relevant information from memory.

This perspective suggests that DK options should be especially likely to attract respondents under the conditions thought to foster satisficing: low ability to optimize, low motivation to do so, or high task difficulty. Consistent with this reasoning, DK filters attract individuals with more limited cognitive skills, as well as those with relatively little knowledge and exposure to information about the attitude object (for a review, see Krosnick 1999). In addition, DK responses are especially common among people for whom an object is low in

personal importance, of little interest, and arouses little affective involvement. This may be because of lowered motivation to optimize under these conditions. Furthermore, people are especially likely to say DK when they feel they lack the ability to formulate informed opinions (i.e., subjective competence), and when they feel there is little value in formulating such opinions (i.e., demand for opinionation). These associations may arise at the time of attitude measurement: low motivation inhibits a person from drawing on knowledge available in memory to formulate and carefully report a substantive opinion of an object.

DK responses are also more likely when questions appear later in a questionnaire, at which point motivation to optimize is presumably waning (Culpepper, Smith, & Krosnick, 1992; Krosnick et al., 2002; Dickinson & Kirzner 1985; Ferber, 1966; Ying, 1989). Also, DK responses become increasingly common as questions become more difficult to understand (Converse, 1976; Klare, 1950).

Hippler and Schwarz (1989) proposed still another reason why DK filters may discourage reporting of real attitudes: Strongly-worded DK filters (e.g., "or haven't you thought enough about this issue to have an opinion?") might suggest that a great deal of knowledge is required to answer a question and thereby intimidate people who feel they might not be able to adequately justify their opinions. Consistent with this reasoning, Hippler and Schwarz found that respondents inferred from the presence and strength of a DK filter that follow-up questioning would be more extensive, would require more knowledge, and would be more difficult. People motivated to avoid extensive questioning or concerned that they could not defend their opinions, might be attracted towards a DK response.

A final reason why people might prefer the DK option to offering meaningful opinions is the desire not to present a socially undesirable or unflattering image of themselves. Consistent

with this claim, many studies found that people who offered DK responses frequently would have provided socially undesirable responses (Cronbach, 1950, p. 15; Fonda, 1951; Johanson, Gips, & Rich, 1993; Kahn & Hadley, 1949; Rosenberg, Izard, & Hollander, 1955).

Taken together, these studies suggest that DKs often result not from genuine lack of opinions but rather from ambivalence, question ambiguity, satisficing, intimidation, and self-protection. In each of these cases, there is something meaningful to be learned from pressing respondents to report their opinions, but DK response options discourage people from doing so. As a result, data quality does not improve when such options are explicitly included in questions.

In order to distinguish "real" opinions from "non-attitudes," follow-up questions that measure attitude strength may be used. Many empirical investigations have confirmed that attitudes vary in strength, and the task respondents presumably face when confronting a "don't know" response option is to decide whether their attitude is sufficiently weak to be best described by that option. But because the appropriate cut point along the strength dimension is both hard to specify and unlikely to be specified uniformly across respondents it seems preferable to encourage people to report their attitude and then describe where it falls along the strength continuum (see Krosnick et al., 1993 and Wegener et al., 1995 for a discussion of the nature and measurement of the various dimensions of strength).

Social Desirability Response Bias

For many survey questions, respondents have no incentive to lie, so there is no reason to believe they intentionally misreport. On questions about socially desirable (or undesirable) matters, however, there are grounds for expecting such misreporting. Theoretical accounts from sociology (Goffman, 1959) and psychology (Schlenker & Weingold, 1989) assert that in pursuing goals in social interaction, people attempt to influence how others see them. Being

viewed more favorably by others is likely to increase rewards and reduce punishments, which may motivate people not only to convey more favorable images of themselves than is warranted, but possibly even to deceive themselves as well (see Paulhus, 1984, 1986, 1991).

The most commonly cited evidence for misreporting in surveys comes from record-check studies, in which respondent answers are compared against entries in official records. Using records as the validation standard, many studies found that more people falsely reported in the socially desirable direction than in the socially undesirable one (Parry and Crossley, 1950; Locander, Sudman and Bradburn, 1976). For example, many more people said they voted when polling place records showed they did not vote than said they did not vote when records showed they did (Traugott and Katosh, 1981).

Errors in records themselves, as well as mistakes made in matching respondents to records, mean that the disparity between records and self-reports is not necessarily due to social desirability bias (see, for example, Presser, Traugott, and Traugott, 1990). However, several other approaches to studying the matter have also found evidence consistent with social desirability bias. One such approach, the "bogus pipeline technique," involves telling people 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). People are more willing to report illicit substance use under these conditions than in conventional circumstances (Evans, Hansen, & Mittlemark, 1977; Murray & Perry, 1987). Likewise, Caucasians are more willing to ascribe undesirable personality characteristics to African-Americans (Sigall & Page, 1971; Pavlos, 1972, 1973) and are more willing to report disliking African-Americans (e.g., Allen, 1975) under bogus pipeline conditions than in conventional ones.

Evidence of social desirability bias also comes from analyses of interviewer effects. The

presumption here is that the observable characteristics of an interviewer may indicate to a respondent the answer the interviewer considers desirable. If respondents reply in a way that corresponds with the interviewers' characteristics, it suggests that the respondents tailored their answers accordingly. Several investigations 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, 1988; Campbell, 1981; Schuman & Converse, 1971) and that white respondents express more favorable attitudes toward African-Americans when responding to African-American interviewers than to white interviewers (Campbell, 1981; Cotter, Cohen, & Coulter, 1982; Finkel, Guterbock, & Borg, 1991).

These findings suggest that eliminating the interviewer (or at least eliminating the interviewer's awareness of the respondent's answer) may reduce social desirability response bias. Consistent with this reasoning, Catholics in one study were more likely to report favoring legalized abortion and birth control on a self-administered questionnaire than to an interviewer (Wiseman, 1972); whites in another study reported more racial prejudice on a self-administered questionnaire than to an interviewer-administered one (Krysan, 1998); and respondents in many studies were more likely to report illicit drug use under self-administration than to interviewers (Tourangeau and Yan, 2007). Similarly, Kreuter, Presser, and Tourangeau (2008) found fewer socially desirable answers using web administration and interactive voice response, both of which eliminate the interviewer, than in a comparable interviewer-administered survey.

Offering anonymity on self-administered questionnaires should further reduce social pressure, and thus may likewise reduce social desirability bias. Paulhus (1984) found that more desirable personality characteristics were reported when people were asked to write their names, addresses and telephone numbers on their questionnaire than when they told not to put

identifying information on the questionnaire (see also Gordon, 1987).

A quite different approach to making answers anonymous involves the "randomized response technique" (Warner, 1965). Here, respondents answer one of various questions, depending upon the outcome of a randomization device. For instance, if a coin flip comes up heads, respondents are asked to answer a nonsensitive item whose distribution is known (e.g., "Were you born in April?") If it comes up tails, they are asked to answer the focal sensitive item (e.g., "Have you ever had an abortion?"). As only the respondent knows the outcome of the randomization device, the researcher does not know which question each person is answering, and thus people may feel freer to be honest. In a meta-analysis of 38 studies, Lensvelt-Mulders, et al. (2005) found that the randomized response technique significantly reduced socially desirable answers. However, most respondents probably do not understand the procedure, which may cause them to not follow the instructions. Edgell, Himmelfarb, and Duchan (1982), for example, found that many respondents would not give the directed response to a question if that response was a socially undesirable one and the question was sufficiently sensitive (see also Holbrook and Krosnick, 2007).

An approach similar to the randomized response technique, but one less likely to arouse respondent suspicion or confusion, is the "item count method" (see, e.g., Droitcour et al., 1991). This approach randomly assigns respondents to one of two lists of items that differ only in whether a focal sensitive item is present. Respondents are asked how many of the items, in total, apply to them, not which apply to them. If the items are chosen appropriately, essentially no one will choose all or none, so it will be possible to estimate the proportion to which the focal item applies without knowing the identity of any particular respondent to whom it applies. As is true for the randomized response technique, however, the item count method introduces an additional

source of sampling error, which means that larger sample sizes are required. Experiments have found that when compared to direct self-reports, the item count method often yielded more reports of socially undesirable behaviors or attitudes (for reviews, see Holbrook and Krosnick, in press; Tourangeau and Yan, 2007). In the instances where this difference did not appear, it could have been because social desirability bias did not distort the direct self-reports.

Another method designed to reduce social desirability bias attempts to save face for respondents by legitimating the less desirable response option. The most common approach involves noting in the question that many people do not engage in the socially desirable behavior, for instance, "In talking to people about elections we often find that a lot of people were not able to vote because they weren't registered, were sick, or just didn't have time." Holbrook and Krosnick (2007) showed that this wording reduces voting reports.

In addition, yes/no response options can be converted into multiple response options, only one of which represents the desirable state, for instance:

1. I did not vote in the November 5th election
2. I thought about voting this time, but didn't
3. I usually vote, but didn't this time
4. I am sure I voted in the November 5th election.

Belli et al. (1999) reported that offering these categories reduced voting reports, though their comparisons simultaneously varied other features as well.

Finally, consistent with our advice in the preceding section on don't knows, it is better not to provide explicit DK options for sensitive items, as they are more apt to provide a cover for socially undesirable responses.

Recall Error

Aside from motivated misreporting due to concern about social desirability, questions about the past are subject to two major sources of error. The first, and most fundamental, is

comprehension. The query "During the last month, have you purchased any household furniture?" for instance, may be compromised by varying interpretations of "household furniture": some people will think lamps count, whereas others will not, and the same will be true for other purchases such as beds. Thus it is critical for questionnaire designers to attend to the ways in which terms are interpreted by respondents (for which pretesting, the subject of our penultimate section, can be invaluable). But this is true for all items, not just those about the past. The second major source of error -- frailties of memory -- is usually of greater concern for retrospective items than for other kinds of items. In this section we review the questionnaire strategies that have been proposed to minimize recall error.[6]

At the outset, it is important to note that recall aids can only be effective for material that was encoded in memory. Although this point seems obvious, surveys sometimes ask for information respondents never knew. Lee et al. (1999), for example, showed that the very low accuracy of parents' reports about their children's immunizations arose because many parents never encoded the information. In these cases, asking respondents to consult records is probably the only way to improve reporting.

When information *is* encoded in memory, its retrieval is strongly affected by both the information's salience and the elapsed time since the information was encoded. Unless the information is both recent and salient it may not come quickly to mind. Thus time spent recalling is often related to accurate reporting. Since respondents may model their behavior on that of the interviewer, instructing interviewers to read more slowly is one way to promote respondent deliberation (and it has the added benefit of making question comprehension easier

[6] The strategies we review generally apply to questions about objective phenomena (typically behavior). For a review of problems associated with the special case of recalling attitudes, see Smith (1984) and Markus (1986).

for respondents).[7] Adding redundant phrases or sentences, thereby lengthening the question, may likewise encourage respondent deliberation by increasing the time available for recall (Laurent, 1972).

Other ways to increase the time taken to answer questions include providing explicit instructions to respondents about the importance of carefully searching their memory (e.g., not saying the first thing that comes to mind); formally asking respondents to commit to doing a good job in line with the instructions; and having the interviewer provide positive feedback to respondents when they appear to be satisfying the instructions. Cannell, Miller, and Oksenberg (1981) showed that these methods, each of which needs to be built into the questionnaire, improved reporting (see also Kessler et al., 2000).

Irrespective of how much time or effort the respondent invests, however, some information will be difficult to recall. When records are available, the simplest approach to improving accuracy is to ask respondents to consult them. Alternatively, respondents may be asked to enter the information in a diary at the time of encoding or shortly thereafter. This requires a panel design in which respondents are contacted at one point and the diaries collected at a later point (with respondents often contacted at an intermediate point to remind them to carry out the task).[8] For discussions of the diary method, see Verbrugge (1980) and Sudman and Ferber (1979).

---

[7] As a reminder to the interviewer of the importance of a slower pace, pause notations may be included in the text of the question, e.g.: "In a moment, I'm going to ask you whether you voted on Tuesday, November 5th (PAUSE) which was ____ days ago. (PAUSE) Before you answer, think of a number of different things that will likely come to mind if you actually did vote this past election day; (PAUSE) things like whether you walked, drove or were driven. (PAUSE) After thinking about it, you may realize that you did not vote in this particular election. (PAUSE) Now that you have thought about it, which of these statements best describes you: I did not vote in the November 5th election; (PAUSE) I thought about voting but didn't; (PAUSE) I usually vote but didn't this time; (PAUSE) I am sure I voted in the November 5th election."

[8] The diary approach – by sensitizing respondents to the relevant information -- may also be used to gather information that respondents would otherwise not encode (e.g., children's' immunizations). But a potential drawback of the method is that it may influence behavior, not just measure it.

Accuracy may also be increased by reducing the burden of the task respondents are asked to perform. This can be done by simplifying the task itself or by assisting the respondent in carrying it out. One common way of simplifying the task is to shorten the reference period. Respondents will have an easier time recalling how often they have seen a physician in the last month than in the last year, and it is easier to recall time spent watching television yesterday than last week.

Most reference periods, however, will be subject to telescoping -- the tendency to remember events as having happened more recently (forward telescoping) or less recently (backward telescoping) than they actually did. Neter and Waksberg (1964) developed the method of bounded recall to reduce this problem. This involves a panel design, in which the second interview asks respondents to report about the period since the first interview (with everything reported in the second interview compared to the reports from the initial interview to eliminate errors). Sudman, Finn, and Lannom (1984) proposed that at least some of the advantages of bounding could be obtained in a single interview, by asking first about an earlier period and then about the more recent period of interest. This was confirmed by an experiment they did, as well as by a similar one by Loftus et al. (1990).

Another way of simplifying the task involves decomposition: dividing a single question into its constituent parts. Cannell et al. (1989), for example, suggested that the item:

> During the past 12 months since July 1st 1987, how many times have you seen or talked with a doctor or a medical assistant about your health?

can be decomposed into four items (each with the same twelve month reference period): overnight hospital stays; other times a doctor was seen; times a doctor was not seen but a nurse or other medical assistant was seen; and times a doctor, nurse or other medical assistant was

consulted by telephone.[9]

In self-administered modes, checklists can sometimes be used to decompose an item. Experimental evidence suggests that checklists should be structured in "yes-no" format as opposed to "check-all-that-apply," partly because respondents take longer to answer forced choice items, and partly because forced choice results are easier to interpret (Smyth et al., 2006).

When it is not feasible to simplify the task, several methods may be used to assist the respondent in carrying it out. All involve attempts to facilitate recall by linking the question to memories related to the focal one. Thus Loftus and Marburger (1983) reported that the use of landmark events (e.g., "since the eruption of Mt. St. Helens…) appeared to produce better reporting than the more conventional approach (e.g., "in the last six months…"). Similarly, Means et al. (1991) and Belli et al. (2007) found that calendars containing key events in the respondent's life improved reporting about other events in the respondent's past.[10]

Another way to aid recall is to include question cues similar to those that were present at the time of encoding. Instead of asking whether a respondent was "assaulted," for instance, the inquiry can mention things the respondent might have experienced as assault -- whether someone used force against the respondent:

> with any weapon: for instance, gun, knife, scissors;
> with anything like a ball bat, frying pan, a chair, or stick;
> by something thrown, such as a rock, bottle, dish, hot liquids;
> by grabbing, punching, choking, scratching, or biting;
> with a sexual attack (Biderman et al., 1986: 92).

This kind of cuing may not only improve recall; it also more clearly conveys the task (by

[9] Belli et al. (2000), however, suggest that decomposition is less good for measuring nondistinctive, frequent events.

[10] As the administration of the calendars in both studies involved conversational or flexible interviewing -- a departure from conventional standardized interviewing -- further research is needed to determine how much of the improved reporting was due to the calendar, per se, and how much to interviewing style.

defining "assault"). But the cues must cover the domain well, as events characterized by uncued features are apt to be underreported relative to those with cued features.[11]

As we noted in the earlier section on open versus closed questions, when asking about amounts, open questions are typically preferable to closed questions, because category ranges using absolute amounts can be interpreted in unwanted ways (Schwarz et al., 1985), and categories using vague quantifiers (e.g., "a few," "some," and "many") can be interpreted differently across respondents (Schaeffer, 1991).

When quantities can be expressed in more than one form, accuracy may be improved by letting respondents select the reporting unit they are most familiar with. In asking about job compensation, for instance, respondents ought to be able to choose whether to report in hourly, weekly, annual, or other terms, as opposed to the researcher choosing a unit for everyone. More generally, given the risk of error, it is usually best to avoid having respondents perform computations that researchers can perform from respondent-provided components.

Question Order

Survey results may be affected not only by the wording of a question, but by the context in which the question is asked. Thus, decisions about the ordering of items in a questionnaire -- fashioning a questionnaire from a set of questions -- should be guided by the same aim that guides wording decisions -- minimizing error.

Question order has two major facets: serial (location in a sequence of items) and semantic (location in a sequence of meanings). Both may affect measurement by influencing the cognitive processes triggered by questions.

[11] Place cues may also aid recall. Thus in the context of crime, one might ask whether victimizations occurred at home, work, school, while shopping, and so on. Likewise, cues to the consequences of events may be helpful. In the case of crime, for example, one might ask respondents to think about times they were fearful or angry (Biderman et al., 1986). On the use of emotions cues, more generally, see Kihlstrom et al. (2000).

Serial Order Effects

Serial order can operate in at least three ways: by affecting motivation, promoting learning, and producing fatigue.

Items at the very beginning of a questionnaire may be especially likely to influence willingness to respond to the survey, because they can shape respondents' understanding of what the survey is about and what responding to it entails. Thus, a questionnaire's initial items should usually bear a strong connection to the topic and purpose that were described in the survey introduction, engage respondent interest, and impose minimal respondent burden. This often translates into a series of closed attitude questions, though factual items can be appropriate as long as the answers are neither difficult to recall nor sensitive in nature. It is partly for this reason that background and demographic characteristics most often come at the end of questionnaires.

Conventional wisdom holds that responses to early items may be more prone to error because rapport has not been fully established or the respondent role has not been completely learned. We know of no experiments demonstrating either of these effects, although Andrews (1984) reported nonexperimental evidence suggesting that questions performed less well at the very beginning of a questionnaire. These considerations support the recommendation that difficult or sensitive items should not be placed early in a questionnaire.

Although respondent learning can be advantageous, it may be disadvantageous in the case of screening items -- those with follow-up questions that are asked only if the original item was answered a particular way (usually "yes"). After learning that answering such questions in a certain way can lengthen the interview, respondents may falsely answer later screening items in order to avoid the contingent questions. Several experiments have yielded evidence suggesting

this happens (Jensen, Watanabe, and Richters, 1999; Lucas et al., 1999; Duan et al., 2007; Kreuter, McCulloch, and Presser, 2009). Although it is possible that the reduction in "yes" answers to later screening items in these experiments was due to improved reporting (because respondents better understand later questions), the weight of the evidence suggests this was not the case.[12] Thus, measurement for multiple screening items is likely to be improved by grouping them together and asking contingent items only after all the screening questions have been administered.[13]

Later items in a questionnaire may also suffer from fatigue effects if respondents become tired. This possibility has been examined in a variety of experiments assessing the impact on data quality of earlier versus later item placement. Consistent with expectations about fatigue and satisficing, several studies have found higher missing data levels, greater agreement, less detailed answers, or less differentiation among items when they appear later in a questionnaire compared to the same items placed earlier (Johnson et al., 1974; Kraut et al., 1975; Herzog and Bachman, 1981; Backor, Golde, and Nie, 2007). Most of the studies reporting such effects involved self-administered questionnaires. Two experiments that found little, if any, difference by item position involved interviewer-administered surveys (Clancy and Wachsler, 1971; Burchell and Marsh, 1992). The possibility that fatigue effects might be slower to set in during interviewer-administered surveys than in self-administered surveys needs to be tested directly in future research.

[12] In a related vein, Peytchev, Couper, McCabe, and Crawford (2006) found that visible skip instructions in the scrolling version of a web survey led more respondents to choose a response that avoided subsequent questions for an item on alcohol use (though not for one on tobacco use) compared to a page version with invisible skips. For findings on related issues, see Gfroerer, Lessler, and Parsley (1997).

[13] Paper and pencil administration constitutes an exception to this rule as the skip patterns entailed by the recommendation are apt to produce significant error in that mode.

Semantic Order Effects

Throughout a questionnaire, items should flow coherently, which usually requires that items on related topics be grouped together.[14] Coherent grouping can facilitate respondents' cognitive processing, e.g., by specifying the meaning of a question more clearly or making retrieval from memory easier. Consistent with this logic, Knowles (1988; see also Knowles and Byers, 1996) found that serial order affected item performance in batteries of personality items. Although order did not influence item means, it did alter item-to-total correlations: the later an item appeared in a unidimensional battery, the more strongly answers to the item correlated with the total score. Put differently, the more questions from the battery an item followed, the more apt it was to be interpreted in the intended manner and/or the more readily respondents retrieved information relevant to the answer. However, Smith (1983) reported inconsistent results on the effects of grouping items, and others (Metzner and Mann, 1953; Baehr, 1953; and Martin, 1980) have found no effect.[15]

A different kind of effect of grouping on retrieval was reported by Cowan, Murphy, and Wiener (1978), who found that respondents reported significantly more criminal victimization when the victimization questions followed a series of attitudinal questions about crime. Answering earlier questions about crime may have made it easier for respondents to recall victimization episodes.

Although grouping related questions may improve measurement, it can lead to poorer assessment under some circumstances. For instance, several experiments have shown that

[14] Although context can affect judgments about whether or not items are related, this effect is likely to be restricted to judgments about items on the same or similar topics.

[15] Couper, Traugott, and Lamias (2001) and Tourangeau, Couper, and Conrad (2004) found that correlations between items in a web survey were a little stronger when the items appeared together on a single screen than when they appeared one item per screen.

respondents' evaluations of their overall life satisfaction were affected by whether the item followed evaluations of specific life domains, but the effect's nature depended on the number of previous related items. When the general item was preceded by a single item about marital satisfaction, some respondents assumed -- having just been asked about their marriage -- that the general item was inquiring about other aspects of their life, so they excluded marital feelings. By contrast, when the general item was preceded by items about several other domains -- including marriage -- then respondents were apt to assume the general item was asking them to summarize across the domains, and thus they were likely to draw on feelings about their marriage in answering it (Schwarz, Strack, and Mai, 1991; Tourangeau, Rasinski, and Bradburn, 1991).[16]

The results from these experiments suggest a qualification of the conventional advice to order related questions in a "funnel," from more general to more specific. Although "general" items are more susceptible to influence from "specific" ones than vice versa (because more general items are more open to diverse interpretation), these context experiments suggest that such influence can improve measurement by exerting control over context (and therefore reduce the diversity of interpretations).

Changing the weights respondents give to the factors relevant to answering a question is another way in which context operates -- by influencing the extent to which a factor is salient or available to the respondent at the time the question is posed. In one of the largest context effects ever observed, many fewer Americans said that the United States should admit communist reporters from other countries when that item was asked first than when it followed an item that asked whether the Soviet Union should admit American reporters (Schuman and Presser, 1981). In this case, a consistency dynamic was evoked when the item came second (making a

[16] Similar findings for general and specific ratings of communities have been reported by Willits and Saltiel (1995).

comparison explicit), but not when it came first (leaving the comparison implicit at best).[17]

In other cases, context can influence the meaning of response options by changing the nature of the standard used to answer a question. For instance, ratings of Bill Clinton might differ depending on whether they immediately follow evaluations of Richard Nixon or of Abraham Lincoln (cf. Carpenter and Blackwood, 1979).

When question ordering affects the meaning of response options or the weighting of factors relevant to answering an item, one context does not necessarily yield better measurement than another. Instead, the effects reflect the fact that choices -- in "real world" settings no less than in surveys -- are often inextricably bound up with the contexts within which the choices are made (Slovic, 1995). Thus decisions about how to order items should be informed by survey aims. When possible, question context should be modeled on the context to which inference will be made. In an election survey, for instance, it makes sense to ask about statewide races after nationwide races, since that is the order in which the choices appear on the ballot. But in the many cases that have no single real-world analog, consideration should be given to randomizing question order.[18]

Although context effects can be unpredictable, they tend to occur almost exclusively among items on the same or closely related topics (Tourangeau, Singer, and Presser, 2003). Likewise the effects are almost always confined to contiguous items (Smith, 1988; but for an

[17] Lorenz, Saltiel, and Hoyt (1995) found similar results for two pairs of items, one member of which asked about the respondents' behavior toward their spouses and the other of which asked about their spouses' behavior toward them.

[18] When the survey goal includes comparison to results from another survey, replicating that survey's questionnaire context is desirable.

exception to this rule, see Schuman, Kalton, and Ludwig, 1983).[19] Schwarz and Bless (1992) and Tourangeau, Rips, and Rasinski (2000) provide good theoretical discussions of survey context. An important tool for identifying potential order effects in a questionnaire is pretesting, to which we turn next.

Testing and Evaluating Questionnaires

No matter how closely a questionnaire follows recommendations based on best practices, it is likely to benefit from pretesting: a formal evaluation carried out before the main survey. This is because best practice recommendations provide little guidance about most specific wording choices or question orderings. In addition, particular populations or measures may pose exceptions to the rules. As a result, questionnaire construction, although informed by science, remains a craft, and pretesting (itself a mix of science and craft) can provide valuable assistance in the process.

Some evaluation methods require administration of the questionnaire to respondents, whereas others do not. Methods not requiring data collection, which are therefore relatively inexpensive to conduct, rely either on human judgment (in some cases by experts, in others by nonexperts) or on computerized judgments. These methods include expert review, forms appraisal, artificial intelligence programs, and statistical modeling. Methods that involve data collection, which are more expensive to carry out, vary along two dimensions: whether they explicitly engage the respondent in the evaluation task -- what Converse and Presser (1986) call participating, as opposed to undisclosed, pretests -- and whether they are conducted in conditions similar to those of the main survey. These methods include cognitive interviews, behavior coding, vignettes, and debriefings of interviewers and/or respondents. For a more detailed

[19] With paper and pencil self-administration and some computerized self-administration, respondents have an opportunity to review later questions before answering earlier ones. Thus, in these modes, later items can affect responses to earlier ones (Schwarz and Hippler, 1995), although such effects are probably not common.

review of pretesting methods, see Presser et al. (2004).[20]

Methods Without Data Collection

Probably the least structured evaluation method is expert review, in which one or more experts critiques the questionnaire. The experts are typically survey methodologists, but they can be supplemented with specialists in the subject matter(s) of the questionnaire. Reviews are done individually or as part of a group discussion.

As many of the judgments made by experts stem from rules, attempts have been made to draw on these rules to fashion an evaluation task that nonexperts can do. Probably the best known of these schemes is the Questionnaire Appraisal System (QAS), a checklist of 26 potential problems (Willis and Lessler 1999; see also Lessler and Forsyth 1996). In an experimental comparison, Rothgeb, Willis, and Forsyth (2001) found that the QAS identified nearly every one of 83 items as producing a problem whereas experts identified only about half the items as problematic -- suggesting the possibility of numerous QAS false positives. In a smaller-scale analysis of 8 income items, by contrast, van der Zouwen and Smit (2004) reported substantial agreement between QAS and expert review.

Evaluations may also be computerized. The Question Understanding Aid (QUAID) -- computer software based partly on computational linguistics -- is designed to identify questions that suffer from five kinds of problems: unfamiliar technical terms, vague or imprecise predicate or relative terms, vague or imprecise noun phrases, complex syntax, and working memory overload (Graesser et al., 2006). Users enter the item text and QUAID compares the words to several databases and performs various calculations. For example, it identifies a word as vague if its concreteness value in a psycholinguistics database is less than a threshold, and it diagnoses

[20] Prior to pretesting, researchers will often benefit from self-administering their questionnaires (role playing the respondent), which provides an opportunity for them to discover the difficulties they have answering their own questions.

a question as involving complex syntax if the number of words before the main verb or main clause exceeds a threshold. The threshold levels were set to maximize the correlations with expert ratings for a small set of items.[21]

A quite different computerized approach predicts an item's measurement properties using an equation developed from a meta-analysis of 87 multi-trait multi-method studies of survey questions (Saris and Gallhofer, 2007). Users of the Survey Quality Predictor (SQP) assign each question a value for each of approximately 50 variables ranging from objective characteristics of the item, such as type of response options, to subjective ones, such as the item's social desirability. The program then outputs coefficients for each item's reliability, validity, and method effect. In their study of 8 income questions, van der Zouwen and Smit (2004) found no association between SQP scores and ratings from either experts or the QAS. But the very small number of items suggests caution in generalizing from these results.[22]

Methods With Data Collection

Unlike methods not involving data collection, which can only make predictions about whether items cause problems, methods using data collection provide evidence of whether the items, in fact, cause problems. The most common form of pretest data collection -- conventional pretesting -- involves administering a questionnaire to a small sample of the relevant population under conditions close to, or identical to, those of the main survey. Interviewers are informed of the pretest's objectives, but respondents are not. The data from conventional pretests consist partly of the distribution of respondent answers to the questions, but mainly of the interviewers' assessments of how the questions worked, which are typically reported at a group debriefing

[21] QUAID may be accessed at http://mnemosyne.csl.psyc.memphis.edu/QUAID/quaidindex.html.

[22] SQP may be accessed at http://www.sqp.nl.

discussion (though sometimes on a standardized form instead of, or in addition to, the group discussion).

Conventional pretest interviews may be used as the foundation for several other testing methods. Behavior Coding, Response Latency, Vignettes, and Respondent Debriefings may all be grafted on to conventional pretest interviews.

Behavior coding measures departures from the prototypical sequence in which the interviewer asks the question exactly as it appears in the questionnaire and then the respondent provides an answer that meets the question's aim. Coding may be carried out by monitors as interviews are conducted or (more reliably) from recordings of the interviews. The most basic code (e.g., Fowler and Cannell, 1996) identifies departures the interviewer makes from the question wording as well as departures the respondent makes from a satisfactory answer, for instance, requesting clarification or expressing uncertainty.[23] Hess, Singer, and Bushery (1999) found that problematic respondent behavior as measured by behavior codes was inversely related to an item's reliability. Dykema, Lepkowski, and Blixt (1997) found that several respondent behavior codes were associated with less-accurate answers (though, for one item, substantive changes in the interviewer's reading of the question were associated with more accurate answers).

Response latency measures the time it takes respondents to answer a question. It may be assessed either during an interview by the interviewer's depressing a key when she finishes asking an item and then again when the respondent begins his answer, or after the interview is completed by listening to recordings (which, as with behavior coding, is less error-prone). Unfortunately, the interpretation of longer times is not always straightforward, as delays in

[23] More elaborate behavior codes (e.g., van der Zouwen and Smit, 2004) consider interaction sequences, e.g., the interviewer reads the question with a minor change, followed by the respondent's request for clarification, which leads the interviewer to repeat the question verbatim, followed by the respondent answering satisfactorily.

responding could mean that a question is difficult to process (usually a bad sign) or that the question encourages thoughtful responding (often a good sign). The one study we know of that addresses this issue with validation data (Draisma and Dijkstra, 2004) found that longer response latencies were associated with more incorrect answers, though another study that addressed the issue more indirectly (Bassili and Scott, 1996) reported mixed results.

Vignettes describe hypothetical situations that respondents are asked to judge. They have been adapted to pretesting to gauge how concepts conveyed in questions are understood. In their test of the meaning of the Current Population Survey's work item, for example, Campanelli, Rothgeb, and Martin (1989) administered vignettes like the following:

> I asked you a question about working last week. Now I'm going to read a list of examples. After each, please tell me whether or not you think the person should be reported as working last week. (1) Last week, Susan only did volunteer work at a local hospital. Do you think she should be reported as working last week? (2) Last week, Amy spent 20 hours at home doing the accounting for her husband's business. She did not receive a paycheck. Do you think she should be reported as working?

The answers shed light on the correspondence between the respondents' conception of work and the definition of work intended by the Current Population Survey. For further details of vignettes, see Martin (2004).

Respondent debriefings refer to the entire class of direct or indirect queries about survey items. Open questions on the order of "Why do you say that?" posed after a respondent has answered an item provide information about what the respondent had in mind when answering the item and thereby can reveal how the item was interpreted. Other debriefing questions focus directly on aspects of the question (e.g., What did you think I meant by...?) or on aspects of choosing an answer (e.g., How difficult was it to answer the question about…?). These inquiries

may be posed immediately after the focal item (Schuman, 1966) or after the entire interview, in which case the questions need to be repeated (Belson, 1981). For a more detailed overview of respondent debriefings, see Martin (2004).

Cognitive interviewing combines many elements of respondent debriefings and produces qualitative data. Respondents are often asked to do two things: (1) think out loud when generating an answer to each question, and (2) answer probes about the questions (e.g., "How would you restate the question in your own words?"). This approach can be valuable for revealing respondent interpretations of a question and identifying misunderstandings that can be prevented by rewording the question. Some researchers have thought that such interviews also reveal the cognitive processes that people implement during actual survey interviews. But in fact, thinking aloud may disrupt such cognitive processes and much of the cognitive processing that yields answers is likely to happen outside of respondent consciousness and would therefore not be revealed by this method (Willis, 2004). For detailed discussions of the method, see Willis (2005) and Beatty and Willis (2007).

Comparisons Across Methods

The multiplicity of testing methods raises questions about their uniqueness -- the extent to which different methods produce different diagnoses. Studies that compare two or more methods applied to a common questionnaire often show a mixed picture -- significant overlap in the problems identified but considerable disagreement as well. The interpretation of these results, however, is complicated by the fact that most of the studies rely on a single trial of each method. Thus differences between methods could be due to unreliability, the tendency of the same method to yield different results across trials.

As might be expected, given its relatively objective nature, behavior coding has been

found to be highly reliable (Presser and Blair, 1994). Conventional pretests, expert reviews, and cognitive interviews, by contrast, have been shown to be less reliable (Presser and Blair, 1994, and DeMaio and Landreth, 2004). The computer methods (QUAID and SQP) may be the most reliable, though we know of no research demonstrating the point. Likewise, the structure of the remaining methods (QAS, response latency, vignettes, and respondent debriefings) suggests their reliability would be between that of conventional pretests, expert reviews and cognitive interviews, on the one hand, and computerized methods, on the other. But, again, we know of no good estimates of these reliabilities.

Inferences from studies that compare testing methods are also affected by the relatively small number of items used in the studies and by the fact that the items are not selected randomly from a well-defined population. Nonetheless, we can generalize to some extent about differences between the methods. The only methods that tend to diagnose interviewer (as opposed to respondent) problems are behavior coding (which explicitly includes a code for interviewer departures from verbatim question delivery) and conventional pretests (which rely on interviewer reports). Among respondent problems, the methods seem to yield many more comprehension difficulties (about the task respondents think the question poses) than performance difficulties (about how respondents do the task), and -- somewhat surprisingly -- this appears most true for cognitive interviews (Presser and Blair, 1994). Conventional testing, behavior coding, QAS, and response latency are also less apt than the other approaches to provide information about how to repair the problems they identify.

Although there is no doubt that all of the methods uncover problems with questions, we know only a little about the degree to which these problems are significant, i.e., affect the survey results. And the few studies that address this issue (by reference to reliability or validity

benchmarks) are generally restricted to a single method, thereby providing no information on the extent to which the methods differ in diagnosing problems that produce important consequences. This is an important area of future research.

Given the present state of knowledge, we believe that questionnaires will often benefit from a multi-method approach to testing. Moreover, when significant changes are made to a questionnaire to repair problems identified by pretesting, it is usually advisable to mount another test to determine whether the revisions have succeeded in their aim and not caused new problems. When time and money permit, this multi-method, multi-iteration approach to pretesting can be usefully enhanced by split sample experiments that compare the performance of different versions of a question or questionnaire (Forsyth, Rothgeb, and Willis, 2004; Schaeffer and Dykema, 2004).

Conclusion

Researchers who compose questionnaires should find useful guidance in the specific recommendations for the wording and organization of survey questionnaires that we have offered in this chapter. In concluding, we offer two more general recommendations. First, questionnaire designers should review questions from earlier surveys before writing their own. This is partly a matter of efficiency – there is little sense in reinventing the wheel – and partly a matter of expertise -- the design of questions and questionnaires involves an art as well as a science and some previous questions are likely to have been crafted by more skillful artisans or those with more resources to develop and test items.

Moreover, even when questions from prior surveys depart from best practice, they may be useful to borrow. This is because replicating questions opens up significant analytical possibilities: Comparisons with the results from other times and from other populations. As such

comparisons require constant wording, it will be appropriate to ask questions that depart from best practice in these cases.

Will such comparisons be affected by the response errors that arise from the departure from best practice? Not if the response errors are constant across the surveys. Unfortunately, most of the literature on question wording and context focuses on univariate effects, and thus we know less about the extent to which response effects vary between groups (i.e., the effect on bivariate or multivariate relationships). Although there is evidence that some response effects (e.g., acquiescence) may affect comparisons between certain groups (e.g., those that differ in educational attainment), there is evidence in other cases for the assumption of "form-resistant correlations" (Schuman and Presser, 1981).

Relevant evidence can be generated by repeating the earlier survey's item on only a random subsample of the new survey, and administering an improved version to the remaining sample. This will not yield definitive evidence (because it relies on the untested assumption that the effect of wording is – or would have been – the same in the different surveys), but it can provide valuable information about the measures.

Second, just as different versions of the "same" item administered to split samples can be instructive, multiple indicators of a single construct (administered to the entire sample) can likewise be valuable. Although the emphasis in the question literature is generally on single items, there is usually no one best way to measure a construct and research will benefit from the inclusion of multiple measures. This is true both in the narrow psychometric sense that error can be reduced by combining measures, as well as in the broader sense of discovery-making when it turns out that the measures do not in fact tap the same construct.

References

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

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.

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods & Research*, *25*, 318–340.

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

Anderson, B. A., Silver, B. D.., & Abramson, P. R. (1988). The effects of the Race of the Interviewer on Race-Related Attitudes of Black Respondents in SRC/CS National Election Studies. *Public Opinion Quarterly, 52*, 289-324.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, *48*, 409-442.

Andrews, F. M. (1990). Some observations on meta-analysis of MTMM studies. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait multimethod studies* (pp. -). Amsterdam, The Netherlands: Royal Netherlands Academy of Arts and Sciences.

Atkinson, R. C., & Shiffrin R. M. (1968). Human memory: a proposed system and its control processes. *The Psychology of Learning and Motivation*, *2*, 89–195.

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

Backor, K., Golde, S., & Nie, N. (2007). *Estimating Survey Fatigue in Time Use Study*. Paper presented at the 2007 International Association for Time Use Research Conference, Washington, D.C.

Baddeley, A. D., & Hitch, G. J. (1977). Recency re-examined. *Attention and Performance*, *6*, 647-665.

Baehr, M. E. (1953). A Simplified Procedure for the Measurement of Employee Attitudes *Journal of Applied Psychology, 37*, 163-167.

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, *60*, 390-399.

Beatty, P. C., & Willis, G. B. (2007). Research Synthesis: The Practice of Cognitive Interviewing. *Public Opinion Quarterly*, *71*, 287-311.

Becker, S. L. (1954). Why an order effect. *Public Opinion Quarterly*, *18*, 271–278.

Belli, R. F, Schwarz, N., Singer, E., & Talarico, J. (2000). Decomposition can harm the accuracy of behavioral frequency reports. *Applied Cognitive Psychology*, *14*, 295–308.

Belli, R. F., Smith, L., M., Andreski, P. M., & Agrawal, S. (2007). Methodological Comparisons Between CATI Event History Calendar and Standardized Conventional Questionnaire Instruments. *Public Opinion Quarterly*, *71*, 603-622.

Belli, R. F., Traugott, M. W., Young, M., & McGonagle, K. A. (1999). Reducing vote overreporting in surveys: social desirability, memory failure, and source monitoring. *Public Opinion Quarterly*, *63*, 90–108.

Belson, W. A. (1981). *The Design and Understanding of Survey Questions*. Aldershot, England: Gower.

Bendig, A. W. (1954). Reliability and the Number of Rating Scale Categories. *Journal of Applied Psychology*, *38*, 38-40.

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

Biderman, A. D., Cantor, D., Lynch, J. P., & Martin, E. (1986). Final Report of the National Crime Survey Redesign Program. Washington, D. C.: Bureau of Social Science Research.

Birkett, N. J. (1986). Selecting the Number of Response Categories for a Likert-type Scale. In Proceedings of the American Statistical Association (pp. 488-492).

Bishop, G. F. (1987). Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly*, *51*, 220–232.

Bishop, G. F. (1990). Issue involvement and response effects in public opinion surveys. *Public Opinion Quarterly*, *54*, 209-218.

Bishop, G. F., Hippler, H. J., Schwarz, N., Strack F. (1988). A comparison of response effects in self-administered and telephone surveys. In, R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), *Telephone Survey Methodology* (pp.321-340). New York: Wiley.

Bishop, G. F., Oldendick, R. W., Tuchfarber, A. J., & Bennett, S. E. (1979). Effects of opinion filtering and opinion floating: Evidence from a secondary analysis. *Political Methodology*, *6*, 293-309.

Bogart L. (1972). *Silent Politics: Polls and the Awareness of Public Opinion*. New York: Wiley.

Burchell, B., & Marsh, C. (1992). The Effect of Questionnaire Length on Survey Response. *Quality and Quantity*, *26*, 233-244.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: the life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*, 197–253.

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

Campanelli, P. C., Rothgeb, J. M., & Martin, E. A. (1989). The Role of Respondent Comprehension and Interviewer Knowledge in CPS Labor Force Classification. In Proceedings of the Section on Survey Research Methods (pp. 425-429). American Statistical Association.

Campbell, A. (1981). *The sense of well-being in America: Recent patterns and trends*. New York: McGraw Hill.

Campbell, D. T., Mohr, P. J. (1950). The effect of ordinal position upon responses to items in a checklist. *Journal of Applied Psychology*, *34*, 62–67.

Cannell, C. F., Miller, P.V., & Oksenberg, L. (1981). Research on interviewing techniques. In *Sociological Methodology*, *11*, 389-437.

Cannell, C. F., Oksenberg, L., Kalton, G., Bischoping, K, & Fowler, F. J. (1989). New Techniques for Pretesting Survey Questions. *Research Report*. Ann Arbor, MI: Survey Research Center, University of Michigan.

Carp, F. M. (1974). Position effects in single trial free recall. *Journal of Gerontology*, *29*, 581-587.

Carpenter, E. H., & Blackwood, L. G. (1979). The Effect of Question Position on Responses to Attitudinal Question. *Rural Sociology, 44*, 56-72.

Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology, 23,* 323-331.

Chan, J. C. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement*, *51*, 531–40.

Cialdini, R. B. (1993). *Influence: Science and Practice* (3rd ed.). New York: Harper Collins.

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

Converse, J. M. (1976). Predicting no opinion in the polls. *Public Opinion Quarterly*, *40*, 515–530.

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 mass publics. In D. E. Apter (Ed.), *Ideology and Discontent* (pp. 206-261). New York: Free Press.

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

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

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

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

Cowan, C. D., Murphy, L. R., & Wiener, J. (1978). Effects of supplemental questions on victimization estimates from the national crime survey. In Proceedings of the Section on Survey Research Methods. American Statistical Association.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*, 3–31.

Culpepper, I. J., Smith, W. R., & Krosnick, J. A. (1992). *The impact of question order on satisficing in surveys*. Paper presented at the Midwestern Psychological Association Annual Meeting, Chicago, IL.

DeMaio, T. J., & Landreth, A. (2004). Do different cognitive interview techniques produce different results? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires*. Hoboken, NJ: Wiley.

Dickinson, J. R., & Kirzner, E. (1985). Questionnaire item omission as a function of within group question position. *Journal of Business Research*, *13*, 71-75.

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

Draisma, S., & Dijkstra, W. (2004). Response Latency and (Para)Linguistic Expressions as Indicators of Response Error. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 131-149). Hoboken, NJ: Wiley.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W. and Ezzati, T. M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In P.P. Bierner, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 185-210). New York: Wiley.

Duan, N., Alegria, M., Canino, G., McGuire, T., & Takeuchi, D. (2007). Survey Conditioning in Self-Reported Mental Health Service Use: Randomized Comparison of Alternative Instrument Formats. *Health Services Research*, *42*, 890-907.

Dykema J., Lepkowski J. M., & Blixt S. (1997). The effect of interviewer and respondent behavior on data quality: Analysis of interaction coding in a validation study. In L. Lyberg, P. Beimer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 287-310). New York: Wiley.

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

Edgell, S. E., Himmelfarb, S., & Duchan, K. L. (1982). Validity of forced response in a randomized response model. *Sociological methods and research, 11,* 89-110.

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

Eurich, A. C. (1931). Four types of examinations compared and evaluated. *Journal of Educational Psychology*, *26*, 268-278.

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

Faulkenberry, G. D., & Mason R. (1978). Characteristics of nonopinion and no opinion response groups. *Public Opinion Quarterly, 42,* 533–43.

Ferber, R. (1966). Item nonresponse in a consumer survey. *Public Opinion Quarterly, 30,* 399–415.

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.

Fonda, C. P. (1951). The nature and meaning of the Rorschach white space response. *Journal of Abnormal Social Psychology*, *46*, 367-377.

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

Forsyth, B. H., Rothgeb, J. M., & Willis, G. (2004). Does Questionnaire Pretesting Make a Difference? An Empirical Test Using a Field Survey Experiment. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.) *Methods for Testing and Evaluating Survey Questionnaires* (pp. 525-546). Hoboken, NJ: Wiley.

Fowler, F. J., Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In N. Schwarz & S. Sudman (Eds.), *Answering Questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 15-36). San Francisco, CA: Jossey-Bass.

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

Gfroerer, J., Lessler, J., & Parsley, T. (1997). Studies of Nonresponse and Measurement Error in the National Household Survey on Drug Abuse. In L. Harrison & A. Hughes (Eds.), *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates* (pp. 273-295). Rockville, MD: National Institute on Drug Abuse.

Gilljam, M., & Granberg, D. (1993). Should we take don't know for an answer? *Public Opinion Quarterly*, *57*, 348-357.

Givon, M. M., & Shapira, Z. (1984). Response to rating scales: a theoretical model and its application to the number of categories problem. *Journal of Marketing Research*, *21*, 410-419.

Glucksberg, S., & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 311–325.

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

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.

Graesser, A. C., Cai, Z., Louwerse, M., & Daniel, F. (2006). Question Understanding Aid. *Public Opinion Quarterly*, 70, 3-22.

Green, P. E., & Rao, V. R. (1970). Rating scales and information recovery – How many scales and response categories to use? *Journal of Marketing*, *34*, 33-39.

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.

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*, 549-559.

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

Hippler, H. J., Schwarz, N. (1989). "No-opinion" filters: a cognitive perspective. *International Journal of Public Opinion Research*, *1*, 77-87.

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

Holbrook, A. L., & Krosnick, J. A. (2005). Do survey respondents intentionally lie and claim that they voted when they did not? New evidence using the list and randomized response techniques. Paper presented at the American Political Science Association Annual Meeting, Washington D. C.

Holbrook, A. L., & Krosnick, J. A. (in press). Social desirability bias in voter turnout reports: Tests using the item count technique. Public Opinion Quarterly.

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

Holmes, C. 1974. A statistical evaluation of rating scales. *Journal of the Marketing Research Society*, *16*, 86-108.

Isard, E. S. (1956). The relationship between item ambiguity and discriminating power in a forced-choice scale. *Journal of Applied Psychology*, *40*, 266-268.

Israel, G. D., & Taylor, C. L. (1990). Can response order bias evaluations? *Evaluation and Program Planning*, *13*, 365-371.

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

Jacoby, J., & Matell, M.S. (1971). Three-point Likert scales are good enough. *Journal of Marketing Research*, *7*, 495-500.

Jenkins, G. D., & Taber, T. D. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology*, *62*, 392-398.

Jensen, P. S., Watanabe, H. K., & Richters, J. E. (1999). Who's up first? Testing for order effects in structured interviews using a counterbalanced experimental design. *Journal of Abnormal Child Psychology*, *27*, 439-436.

Johanson, G. A., Gips, C. J., & Rich, C. E. (1993). If you can't say something nice – A variation on the social desirability response set. *Evaluation. Review*, *17*, 116-122.

Johnson, J. D. (1981). Effects of the order of presentation of evaluative dimensions for bipolar scales in four societies. *Journal of Social Psychology*, *113*, 21-27.

Johnson, W. R., Sieveking, N. A., & Clanton, E. S. (1974). Effects of alternative positioning of open-ended questions in multiple-choice questionnaires. *Journal of Applied Psychology*, *6*, 776-778.

Juster, F. T., & Smith, J. P. (1997). Improving the quality of economic data: lessons from the HRS and AHEAD. *Journal of the American Statistical Association*, *92*, 1268-1278.

Kahn, D. F., & Hadley, J. M. (1949). Factors related to life insurance selling. *Journal of Applied Psychology*, *33*, 132-140.

Kalton, G., Collins, M., & Brook, L. (1978). Experiments in wording opinion questions. *Applied Statistics*, *27*, 149-161.

Kalton, G., Roberts, J., & Holt, D. (1980). The effects of offering a middle response option with opinion questions. *Statistician*, *29*, 65-78.

Katosh, J. P., & Traugott, M. W. (1981). The Consequences of Validated and Self-Reported Voting Measures. *Public Opinion Quarterly*, *45*, 519–535.

Kessler, R. C., Wittchen, H. U., Abelson, J. M., & Zhao, S. (2000). Methodological issues in assessing psychiatric disorder with self-reports. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 229-255). Mahwah, NJ: Lawrence Erlbaum Associates.

Kihlstrom, J. F., Mulvaney, S., Tobias, B. A., & Tobis, I. P. (2000). The emotional unconscious. In E. Eich, J. F. Kihlstrom, G. H. Bower, J. P. Forgas, & P. M. Niedenthal (Eds.), *Cognition and emotion* (pp. 30-86). New York: Oxford University Press.

Klare, G. R. (1950). Understandability and indefinite answers to public opinion questions. *International Journal of Opinion and Attitude Research*, *4*, 91-96.

Klayman, J. & Ha, Y. (1984). *Confirmation, disconfirmation, and information in hypothesis-testing*. Unpublished manuscript, Graduate School of Business, Center for Decision Research, University of Chicago, IL.

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

Klopfer, F. J., & Madden, T. M. (1980). The middlemost choice on attitude items: ambivalence, neutrality, or uncertainty. *Personality and Social Psychology Bulletin*, *6*, 97-101.

Knowles, E. E., & Byers, B. (1996). Reliability shifts in measurement reactivity: Driven by content engagement or self-engagement? *Journal of Personality and Social Psychology*, *70*, 1080-1090.

Knowles, E. S. (1988). Item context effects on personality scales: measuring changes the measure. *Journal of Personality and Social Psychology*, *55*, 312-320.

Komorita, S. S. (1963). Attitude context, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology*, *61*, 327-334.

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement¸ 25*, 987-995.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107-118.

Kraut, A. I., Wolfson, A. D., & Rothenberg, A. (1975). Some effects of position on opinion survey items. *Journal of Applied Psychology*, *60*, 774-776.

Kreuter, F., McCulloch, S. Presser, S. (2009). Filter Questions in Interleafed versus

Grouped Format: Effects on Respondents and Interviewers. Unpublished MS.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly*, *72,* 847-865.

Krosnick, J. A. (1990). American's perceptions of presidential candidates: A test of the projection hypothesis. *Journal of Social Issues, 46,* 159–182.

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. (1992). The impact of cognitive sophistication and attitude importance on response order effects and question order effects. In N. Schwarz & S. Sudman (Eds.), *Order Effects in Social and Psychological Research* (pp. 203-218). New York: Springer.

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology*, *50*, 537-567.

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., & 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., & Fabrigar, L. R. (forthcoming). The handbook of questionnaire design. New York: Oxford University Press.

Krosnick, J. A., & Schuman, H. (1988). Attitude intensity, importance, and certainty and susceptibility to response effects. *Journal of Personality and Social Psychology*, *54*, 940-952.

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

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., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Ruud, P. A., Smith, V. K., Moody, W. R., Green, M. C., & Conaway, M. (2002). The impact of 'no opinion' response options on data quality: non-attitude reduction or invitation to satisfice? *Public Opinion Quarterly*, *66*, 371-403.

Krysan, M. (1998). Privacy and the Expression of White Racial Attitudes. *Public Opinion Quarterly*, *62*, 506-544.

Kuncel, R. B. (1973). Response process and relative location of subject and item. *Educational and Psychological Measurement*, *33*, 545-563.

Kuncel, R. B. (1977). The subject-item interaction in itemmetric research. *Educational and Psychological Measurement*, *37*, 665-678.

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*, 229-242.

Laurent, A. (1972). Effects of question length on reporting behavior in the survey interview. *Journal of the American Statistical Association*, *67*, 298-305.

Lee, L., Brittingham, A., Tourangeau, R., Willis, G., Ching, P., Jobe, J., & Black, S. 1999. Are reporting errors due to encoding limitations or retrieval failure? *Applied Cognitive Psychology, 13*, 43–63.

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

Lehmann, D. R., & Hulbert, J. (1972). Are three-point scales always good enough? *Journal of Marketing Research*, *9*, 444-446.

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

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. (2005). Meta-analysis of randomized response research, thirty-five years of validation, *Sociological Methods & Research*, *33*, 319-348.

Lessler, J. T., & Forsyth, B. H. (1996). A coding system for appraising questionnaires. In N. Schwartz & S. Sudman (Eds.), *Answering questions* (pp. 259-292). San Francisco: Jossey-Bass.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1-55.

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

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology*, *60*, 10-13.

Locander, W., Sudman, S., & Bradburn, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association*, *71*, 269-275.

Loftus, E. F., & Marburger, W. (1983). Since the eruption of Mt. St. Helens, has anyone beaten you up? *Social Cognition*, *11*, 114-120.

Loftus, E. F., Klinger, M. R., Smith, K. D., & Fiedler, J. A Tale of Two Questions: Benefits of Asking More than One Question. *Public Opinion Quarterly*, *54*, 330-345.

Lorenz, F., Saltiel, J., & Hoyt, D. (1995). Question order and fair play: Evidence of even-handedness in rural surveys. *Rural Sociology, 60,* 641–653.

Lucas, C. P., Fisher, P., Piacentini, J., Zhang., H., Jensen, P. S., Shaffer, D., Dulcan, M., Schwab-Stone, M., Regier, D., & Canino, G. (1999). Features of Interview Questions Associated with Attenuation of Symptom Reports. *Journal of Abnormal Child Psychology, 27*, 429-437.

Markus, G. B. (1986). Stability and change in political attitudes: Observed, recalled, and "explained". *Political Behavior*, *8*, 21-44.

Martin, E. (1980). The Effects of Item Contiguity and Probing on Measures of Anomia. *Social Psychology Quarterly*, *43*, 116-120.

Martin, E. (2004). Vignettes and Respondent Debriefing for Questionnaire Design and Evaluation. In S. Presser, J. M. Rothgeb, M. P. Couper, J. L. Lessler, E. Martin, J. Martin & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 149-172). New York: Wiley.

Martin, W. S. (1973). The effects of scaling on the correlation coefficient: A test of validity: *Journal of Marketing Research*, *10*, 316-318.

Martin, W. S. (1978). Effects of scaling on the correlation coefficient: Additional considerations. *Journal of Marketing Research*, *15*, 304-308.

Masters, J. R. (1974). The relationship between number of response categories and reliability of likert-type questionnaires. *Journal of Educational Measurement*, *11*, 49-53.

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.

Matell, M. S., & 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*, 506-509.

Mathews, C. O. (1927). The effect of position of printed response words upon children's answers to questions in two-response types of tests. *Journal of Educational Psychology*, *18*, 445-457.

McClendon, M. J. (1986). Response-order effects for dichotomous questions. *Social Science Quarterly*, *67*, 205-211.

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.

McKelvie, S. J. (1978). Graphic rating scales-How many categories. *British Journal of Psychology*, *69*, 185-202.

Means, B., Swan, G. E., Jobe, J. B., & Esposito, J. L. An alternative approach to obtaining personal history data. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 127-144). New York: Wiley.

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

Metzner, H., & Mann, F. (1953). Effects of grouping related questions in questionnaires. *Public Opinion Quarterly*, *17*, 136-141.

Miller, N., & Campbell, D. T. (1959). Recency and primacy in persuasion as a function of the timing of speeches and measurement. *Journal of Abnormal Social Psychology*, *59*, 1-9.

Miller, W. E. (1982). *American national election study, 1980: Pre and post election surveys*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Mingay, D. J., & Greenwell, M. T. (1989). Memory bias and response-order effects. *Journal of Official Statistics*, *5*, 253-263.

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.

Mondak, J. J. (2001). Developing Valid Knowledge Scales. *American Journal of Political Science*, *45*, 224-238.

Morin, R. (1993, December 6-12). Ask and you might deceive: The wording of presidential approval questions might be producing skewed results. *The Washington Post National Weekly Edition*. p. 37.

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.

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

Neter, J. & Waksberg, J. (1964). A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association*, *59*, 18-55.

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

Norman, D. A. (1973). Memory, knowledge, and the answering of questions. In R. L. Solso (Ed.), *Contemporary issues in cognitive psychology: The Loyola Symposium*. Washington, D. C.: Winston.

O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (1999, May). *Middle alternatives, acquiescence, and the quality of questionnaire data*. Paper presented at the American Association for Public Opinion Research Annual Meeting, St. Petersburg, FL.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Ostrom, T. M., & Gannon, K. M. (1996). Exemplar generation: Assessing how respondents give meaning to rating scales. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 293-441). San Francisco: Jossey-Bass.

Parry, H. J., & Crossley, H. M. (1950). Validity of Responses to Survey Questions. *Public Opinion Quarterly*, *14*, 61-80.

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. Volume 1 in Measures of Social Psychological attitudes Series*. San Diego, CA: Academic Press.

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

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, J. D. (1971). The effects of reversing the order of verbal rating scales in a postal survey. *Journal of the Marketing Research Society*, *14*, 30–44.

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

Payne, S. L. (1950). Thoughts about meaningless questions. *Public Opinion Quarterly*, *14*, 687-696.

Peytchev, A., Couper, M. P., McCabe, S. E., & Crawford, S. D. (2006). Web Survey design. Paging Versus scrolling. *Public Opinion Quarterly*, *70*, 596-607.

Poe, G. S., Seeman, I., McLaughlin, J., Mehl, E., & Dietz, M. (1988). Don't know boxes in factual questions in a mail questionnaire. *Public Opinion Quarterly*, *52*, 212-222.

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

Presser, S., Traugott, M. W., & Traugott, S. (1990). *Vote "over" reporting in surveys: the records or the respondents?* Presented at the International Conference on Measurement Errors, Tucson, AZ.

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., & Singer, E. (Eds.). (2004). *Methods for Testing and Evaluating Survey Questionnaires*. New York: Wiley.

Quinn, S. B., & Belson, W. A. (1969). *The Effects of Reversing the Order of Presentation of Verbal Rating Scales in Survey Interviews*. London: Survey Research Center.

Ramsay, J. O. (1973). The effect of number categories in rating scales on precision of estimation of scale values. *Psychometrika*, *38*, 513-532.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1999). *Measures of political attitudes*. San Diego, CA: Academic Press.

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

Rosenberg, N., Izard, C. E., & Hollander, E. P. (1955). Middle category response: reliability and relationship to personality and intelligence variables. *Educational and Psychological Measurement*, *15*, 281-290.

Rosenstone, S. J., Hansen, J. M., & Kinder, D. R. (1986). Measuring change in personal economic well-being. *Public Opinion Quarterly*, *50*, 176-192.

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

Rothgeb, J., Willis, G., & Forsyth, B. H. (2001). Questionnaire pretesting methods: do different techniques and different organizations produce similar results? In Annual Meeting of the American Statistical Association.

Rubin, H. K. (1940). *A constant error in the Seashore test of pitch discrimination*. Unpublished master's thesis, University of Wisconsin, Madison, WI.

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.

Rundquist, E. A., & Sletto, R. F. (1936) *Personality in the depression*. Minneapolis: University of Minnesota Press.

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

Saris, W. E., & Krosnick, J. A. (2000, May). *The damaging effect of acquiescence response bias on answers to agree/disagree questions*. Paper presented at the American Association for Public Opinion Research Annual Meeting, Portland, OR.

Schaeffer, N. C. (1991). Hardly ever or constantly? Group comparisons and vague quantifiers. *Public Opinion Quarterly*, *55*, 395-423.

Schaeffer, N. C., & Bradburn, N. M. (1989). Respondent behavior in magnitude estimation. *Journal of the American Statistical Association*, *84*, 402-413.

Schaeffer, N. C., & Dykema, J. (2004). A multiple-method approach to improving the clarity of closely related concepts: Distinguishing legal and physical custody of children. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 475-502). New York: Wiley.

Scherpenzeel, A. (1995). Meta-analysis of a European comparative study. In W. E. Saris & A. Munnich (Eds.), *The multitrait-multimethod approach to evaluate measurement instruments* (pp. 225-242). Budapest, Hungary: Eotvos University Press.

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

Schuman, H. (1966). The random probe: a technique for evaluating the validity of closed questions. *American Sociological Review*, *31*, 218–222.

Schuman, H. (1972). Two Sources of Anti-War Sentiment in America. *American Journal of Sociology*, *78*, 513-536.

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: Experiments on question form, wording and context*. New York: Academic Press.

Schuman, H., & Scott, J. (1987). Problems in the use of survey questions to measure public opinion. *Science*, *236*, 957-959.

Schuman, H., Kalton, G., & Ludwig, J. (1983). Context and Contiguity in Survey Questionnaires. *Public Opinion Quarterly*, *47*, 112-115.

Schwarz, N., & Bless, H. (1992). Constructing reality and its alternatives: An inclusion/exclusion model of assimilation and contrast effects in social judgment. In L. L. Martin & A. Tesser (Eds.) *The construction of social judgment* (pp. 217-245). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schwarz, N., & Hippler, H. J. (1991). *Response alternatives: the impact of their choice and presentation order.* In P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Error in Surveys* (pp. 41–56). New York: Wiley.

Schwarz, N., & Hippler, H. J. (1995). Subsequent questions may influence answers to preceding questions in mail surveys. *Public Opinion Quarterly*, *59*, 93-97.

Schwarz, N., & Strack, F. (1985). Cognitive and affective processes in judgments of subjective well-being: a preliminary model. In H. Brandstatter & E. Kirchler (Eds.), *Economic Psychology* (pp. 439-447). Linz, Austria: Tauner.

Schwarz, N., & Wyer, R. S. (1985). Effects of rank-ordering stimuli on magnitude ratings of these and other stimuli. *Journal of Experimental Social Psychology*, *21*, 30-46.

Schwarz, N., Hippler, H. J., & Noelle-Neumann, E. (1992). A cognitive model of response-order effects in survey measurement. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 187-201). New York: Springer-Verlag.

Schwarz, N., Strack, F., & Mai, H. (1991). Assimilation and Contrast Effects in Part-Whole Question Sequences: A Conversational Logic Analysis. *Public Opinion Quarterly*, *55*, 3-23.

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

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

Sherif, C. W., Sherif, M., & Nebergall, R. E. (1965). *Attitude and social change*. Philadelphia: Saunders.

Sherif, M., & Hovland, C. I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change.* New Haven, CT: Yale University Press.

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

Simon, H. A. (1957). *Models of Man*. New York: Wiley.

Slovic, P., (1995). The construction of preference. *American Psychology*, *50*, 364-371.

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

Smith, T. W. (1984). Non-attitudes: A review and evaluation. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (pp. 215-255). New York: Russell Sage.

Smith, T. W. (1988). Context Effects in the General Social Survey. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 57-72). New York: Wiley.

Smith, T. W. (1994a). A comparison of two confidence scales. *GSS Methodological Report No. 80*. Chicago, IL: National Opinion Research Center.

Smith, T. W., & Peterson, B. L. (1985, August). *The impact of number of response categories on inter-item associations: Experimental and simulated results*. Paper presented at the American Sociological Association Meeting, Washington, D. C.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly, 70*, 66-77.

Srinivasan, V., & Basu, A. K. (1989). The metric quality of ordered categorical data. *Marketing Science*, *8*, 205-230.

Stember, H. & Hyman, H. (1949/1950). How interviewer effects operate through question form. *International Journal of Opinion and Attitude Research*, *3*, 493-512.

Sudman, S., & Ferber, R. (1979). *Consumer Panels*. Chicago: American Marketing Association.

Sudman, S., Finn, A., & Lannom, L. (1984). The Use of Bounded Recall Procedures in Single Interviews. *Public Opinion Quarterly*, *48*, 520-524.

Sussman, B. (1978). President's popularity in the polls is distorted by rating questions. *The Washington Post*, pp.

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*, *3*, 299-314.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*, 859-883.

Tourangeau, R., Rasinski, K. A., & Bradburn, N. (1991). Measuring happiness in surveys: A test of the subtraction hypothesis. *Public Opinion Quarterly, 55*, 255-266.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The Psychology of Survey Response*. New York: Cambridge University Press.

Tourangeau, R., Singer, E., & Presser, S. (2003). Context effects in attitude surveys: Effects of remote items and impact on predictive validity. *Sociological Methods & Research*, *31*, 486-513.

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

Vaillancourt, P. M. (1973). Stability of children's survey responses. *Public Opinion Quarterly*, *37*, 373-387.

van der Zouwen, J., & Smit, J. H. (2004). Evaluating survey questions by analyzing patterns of behavior codes and question-answer sequences: A diagnostic approach. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 109-130). Hoboken, NJ: Wiley.

Verbrugge, L. M. (1980). Health diaries. *Medical Care, 18,* 73.

Visser, P. S., Krosnick, J. A., Marquette, J. F., & Curtin, M. F. (2000). Improving election forecasting: Allocation of undecided respondents, identification of likely voters, and response order effects. In P. L. Lavrakas & M. Traugott (Eds.), *Election polls, the news media, and democracy* (pp. 224-260). New York: Chatham House.

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

Warr, P., Barter, J., & Brownridge, G. (1983). On the interdependence of positive and negative affect. *Journal of Personality and Social Psychology*, *44*, 644-651.

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.

Watson, D. (1988). The vicissitudes of mood measurement: Effects of varying descriptors, time frames, and response formats on measures of positive and negative affect. *Journal of Personality and Social Psychology*, *55*, 128-141.

Watson, D. R., & Crawford, C. C. (1930). Four types of tests. *The High School Teacher*, *6*, 282-283.

Wedell, D. H., & Parducci, A. (1988). The category effect in social judgment: Experimental ratings of happiness. *Journal of Personality and Social Psychology*, *55*, 341-356.

Wedell, D. H., Parducci, A., & Geiselman, R. E. (1987). A formal analysis of ratings of physical attractiveness: Successive contrast and simultaneous assimilation. *Journal of Experimental Social Psychology*, *23*, 230-249.

Wedell, D. H., Parducci, A., & Lane, M. (1990). Reducing the dependence of clinical judgment on the immediate context: Effects of number of categories and types of anchors. *Journal of Personality and Social Psychology*, *58*, 319-329.

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: Lawrence Erlbaum Associates.

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

Willis, G. B. (2004). Cognitive interviewing revisited: a useful technique, in theory? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 23-43). Hoboken, NJ: Wiley.

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

Willis, G. B., & Lessler, J. (1999). *The BRFSS-QAS: A guide for systematically evaluating survey question wording*. Rockville, MD: Research Triangle Institute.

Willits, F. K., & Saltiel, J. (1995). Question order effects on subjective measures of quality of life: A two-state analysis. *Rural Sociology*, *57*, 654-665.

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

Ying, Y. (1989). Nonresponse on the center for epidemiological studies–depression scale in Chinese Americans. *International Journal of Social Psychiatry*, *35*, 156-163.

Yzerbyt, V. Y., & Leyens, J. (1991). Requesting information to form an impression: The influence of valence and confirmatory status. *Journal of Experimental Social Psychology*, *27*, 337–356.