Several theoretical hypotheses are developed concerning the relation of question and respondent characteristics to the reliability of survey attitude measurement. To test these hypotheses, reliability is estimated for 96 survey attitude measures using data from five, 3-wave national reinterview surveys—three Michigan Election Panel Surveys and two reinterview studies conducted by the General Social Survey. As hypothesized, a number of attributes of questions are linked to estimated reliability. Attitude questions with more response options tended to have higher reliabilities, although there are some important exceptions. More extensive verbal labeling of numbered response options was found to be associated with higher reliability, but questions explicitly offering a "don't know" alternative were not found to be more reliable. Question characteristics were confounded to an unknown degree with topic differences of questions, which were significantly linked to reliability, leaving the influence of question characteristics on reliability somewhat ambiguous. Characteristics of respondents were also found to be related to levels of reliability. Older respondents and those with less schooling provided the least reliable attitude reports. These results are discussed within a general framework for the consideration of survey errors and their sources.

The Reliability of Survey Attitude Measurement

The Influence of Question and Respondent Attributes

DUANE F. ALWIN
University of Michigan
JON A. KROSNICK
Ohio State University

An attitude is a latent, unobserved predisposition to respond along a positive or negative dimension (e.g. approval vs. disapproval, approach vs. avoidance, satisfaction vs. dissatisfaction etc.) toward an attitude object. The study of the reliability of attitude

AUTHORS' NOTE: *An earlier version of this article was presented at the 1989 meetings of the American Association for Public Opinion Research. The research reported here was supported by Grant R01-AG04743-04 from the National Institute on Aging. We acknowledge the Inter-University Consortium for Political and Social Research for providing access to the National Election Study data and Tom Smith for access to the General*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 20, No. 1, August 1991 139-181

measurement is important because of the ubiquitous nature of the attitude concept in modern social science, because of the pervasive presence of attitude measures in survey research, and because of the difficulties presented in measuring a construct defined as a latent variable (see Alwin 1973). Some researchers have concluded that there is little evidence that stable, underlying attitudes can be said to exist (e.g. Abelson 1972; Wicker 1969), but even among those who accept the theoretical legitimacy of the attitude concept, there is considerable skepticism that the concept applies to all members of the population (Converse 1964, 1970). And, given the typical assumption that attitudes are more difficult to measure than so-called factual material (e.g. Kalton and Schuman, 1982), it is important to focus attention on the reliability with which attitudes can be measured.

In this article we examine several factors contributing to the reliability of attitude measurement in sample surveys. We depart from the view that reliability is a function primarily of the instrument of measurement (e.g. Wiley and Wiley 1970; Achen 1975), arguing instead that reliability is a function of a number of factors, including: (a) the characteristics of the populations of interest, (b) the topics assessed by the question (e.g. facts vs. attitudes, or the type of attitude), (c) the design of the questions, including their wording and context, as well as the response formats provided, and (d) a range of factors affecting the specific conditions of measurement, such as the observational design, the mode of administering the questionnaire, the training of interviewers, and more generally the social situation in which the survey interview is obtained (see Alwin 1989).

Here we focus on two broad sets of factors: the nature of the questions, specifically the attitude objects addressed and characteristics of the response formats; and the characteristics of respondents tied to differences in cognitive capacities and motivation, specifically their level of schooling and their age. We introduce a theoretical framework

---

Social Survey reinterview data. We also gratefully acknowledge the assistance of Frank Mierzwa Merilynn Dielman in data management and analysis, and Evelyn Caviani in manuscript preparation. We also wish to thank Frank Andrews, John Bynner, James Davis, McKee McClendon, Willard Rodgers, Willem Saris, Jacqueline Scott, Tom Smith, and Joseph Woelfel for suggestions on earlier drafts of this material.

for understanding the sources of random errors in survey attitude measurement, and we test several hypotheses based on this theory. For these purposes we estimate reliability for 96 attitude measures from five national panel studies.

THE CONCEPT OF RELIABILITY IN SURVEY RESEARCH

There are a number of misconceptions about the concept of reliability among survey practitioners. It is often suggested, for example, that the correlation or some other measure of association (such as the percentage of agreement or bivariate agreement coefficient) between two measures of the same thing provides an assessment of reliability (see, e.g., Schuman and Presser 1981). This may be true in some instances, but in general it is not. Reliability, as will be explained in greater detail below, refers to the correlational linkage between observed variables and some conception of a "true" variable, and only under certain conditions of design would correlations among survey measures be expected to provide an estimate of reliability.

In order to properly assess reliability one needs, first, a model that specifies the linkage between true and observed variables, second, a research design that permits the estimation of parameters of such a model, and, third, an interpretation of these parameters that is consistent with the concept of reliability. (See Alwin [1989] for a detailed discussion of the problems of estimating the reliability of survey data.)

We use the psychometric concept of reliability, derived from classical true score theory, to assess the extent of random errors in survey responses (Lord and Novick 1968). The psychometric conception of reliability differs from other, perhaps more popular usages of the term. In industry, for example, the term is often used to refer to the absence of "inadvertent, unintentional human actions" that "exceed some limit of acceptability or appropriateness in work performance" (Miller and Swain 1987, pp. 220-21). The term is frequently used in social research to refer to the absolute agreement between measures or codes (e.g. Krippendorff 1970).

The psychometric definition of reliability is both more and less restrictive than these other conceptions. It refers essentially to cor-

relational consistency of response, independent of true individual change. Thus it is limited to random errors, rather than to all such errors, and in this sense is more restrictive than other conceptions of measurement precision. On the other hand, it is less restrictive than ideas of reliability that rely on the idea of "absolute" consistency or agreement (e.g. Smith 1980), in that psychometric reliability theory requires neither a zero intercept in the regression of true scores of multiple measures, nor the identical scaling of the two measures. Reliability, thus, refers to the normed linear relationship between two attempts to measure the same thing, net of true change (Lord and Novick 1968).[1]

According to classical true score theory, an observed score is a function of a true score and a random error score, that is, $y = \tau + \varepsilon$, in a population of individuals for whom the random error model holds (Lord and Novick 1968, pp. 32-34). Under conditions of random error in the measures, the covariance between two or more attempts to measure the same thing reflects true score variance, whereas the variance of replicate measures contain both true variance and random error variance. Reliability is defined as the squared correlation between the observed and true scores, $\rho^2_{y\tau}$, which is equal to the ratio of true-score variance to observed score variance, σ^2_τ/σ^2_y. Thus, in the simplest case of tau-equivalent measures (see Alwin and Jackson 1979), the reliabilities of two survey measures of the same thing are defined as σ_{ij}/σ^2_i and σ_{ij}/σ^2_j for measures y_i and y_j (where σ_{ij} is the covariance between the two measures, and σ^2_i and σ^2_j are their variances).

Because these are estimated population quantities, reliability is clearly a characteristic of a population of persons. Moreover, it is also the case that the amount of measurement error may be affected by the measuring instrument, that is, by those aspects of data collection that depend not on the population being measured, but on the characteristics of survey questions. Of course, it is also conceivable that population characteristics and instrument characteristics can interact in generating random errors, but here we concentrate on the average additive effects of question characteristics independent of population characteristics, and vice-versa.

THE IMPORTANCE OF ESTIMATING RELIABILITY

There are several important reasons for estimating the reliability of single survey items.[2] First, random measurement errors inflate the overall estimated response variance, thus it is important to have knowledge of the extent of such errors (Bohrnstedt 1983). Second, one consequence of this inflation of response variance is that derived estimates of sampling errors are biased (Cleary, Linn, and Walster 1970). Such estimates can be improved by taking into account random measurement errors. Third, another consequence of random measurement errors is that estimates of bivariate relationships are attenuated, and multivariate regression coefficients are biased (Bohrnstedt and Carter 1971). Estimates of the relative importance of variables can be improved by taking into account such errors. Fourth, the study of reliability can focus on the factors that contribute to it, so that random errors of measurement may be reduced. And fifth, given that the reliability of composite scores is a direct function of the reliabilities of the component items, and given that composite scores based on multiple survey items are frequently, although not exclusively, used by attitude researchers, it is important to understand the sources of unreliability in responses to individual items.

Perhaps the most important reason for studying reliability of measurement, which is implicit in several of the points listed above, is its relationship to measurement validity. At the same time, this is one area—the relationship between reliability and validity—where considerable confusion exists. It can be shown that the reliability of measuring of a particular variable sets an upper limit on the magnitude of observed correlations with other variables (Lord and Novick 1968, p. 161). Thus reliability is a necessary condition for empirical validity, in that the correlation of a given variable with another cannot exceed the index of reliability of either variable.[3] It is in this sense that the reliability of responses to a particular survey question sets an upper limit on the magnitude of observed correlations with other variables measured in the survey. It is illogical, however, to reverse the implication. That is, it would be a grave error to infer validity from reliability (see Alwin 1989, p. 283).

SURVEY ATTITUDE MEASUREMENT

Our inferences about the influences of respondent and question characteristics on the reliability of attitude measures are based on the assumption that attitudes fall along a single, latent continuum that ranges from positive to negative. This is compatible both with the way in which attitudes are defined and the ways they are measured. Various aspects of the response categories presumed to represent those latent continua might influence the extent of random error in responses and may influence the use of these categories to map underlying attitudes.

SOURCES OF RANDOM ERRORS IN SURVEY ATTITUDE REPORTS

We consider three sources of random measurement error in survey reports of attitudes: (a) the nonexistence of attitudes, (b) ambiguity in internal attitudinal cues, and (c) ambiguity of response scale alternatives.

Nonattitudes

Respondents who have no firm attitude or opinion on a given public issue, either because they have little knowledge of the issues or have not thought about it, may behave randomly when responding to a survey attitude question. In their study of question wording, Rugg and Cantril (1944, pp. 48-49) acknowledged that response errors were often less likely if persons "had standards of judgement resulting from stable frames of reference," as opposed to situations "where people lack *reliable* standards of judgement and *consistent* frames of reference" (emphasis added).

In an important article, Converse (1964) proposed that some respondents feel pressure during survey interviews to offer opinions in response to attitude questions, when in fact they have none. This is the case, he argued, because respondents assume that interviewers expect them to offer opinions and because opinionated people are presumed to be respected more than persons without many opinions. Because respondents wish to conform to interviewer expectations and to project positive self-images, Converse claimed, they frequently concoct attitude reports during interviews. Such responses are, he said, essen-

tially random choices from among the offered response alternatives. If some respondents do select answers randomly when they lack attitudes, their behavior would increase the amount of random variation in attitude reports.

Critics of Converse's thesis regarding the prevalence of nonattitudes and the potential for random reporting have focused on the sources of random error residing in the questions themselves rather than the respondents. Achen (1975, p. 1229), for example, looking at some of the panel data to be considered below, concludes: "Measurement error is primarily a fault of the instruments not the respondents."[4] The vagueness of attitude questions are often blamed, but it is possible to imagine a number of instrument-related characteristics of survey questions (see below). In any event, these explanations are not mutually exclusive—presumably both may be happening.

Ambiguity in Internal Attitudinal Cues

Some random variation in attitude reports is likely to result from ambiguity in respondents' attitudes. It seems likely that some attitudes are associated with univocal, unambiguous internal cues that come to mind quickly and effortlessly when a person simply thinks about an attitude object. Other attitudes are associated with ambiguous or conflicting internal cues or with cues that are relatively inaccessible and come to mind only as the result of some cognitive effort (Fazio, Herr, and Olney 1984).

This is consistent with Bem's (1972) self-perception theory, which suggests that the internal cues indicating attitudes are sometimes ambiguous. With regard to some issues, some persons have relatively consistent internal cues indicating their attitudes toward the object, whereas others have highly conflicting or ambiguous cues. If this is the case, forcing respondents to choose a single point on an attitude continuum may cause them to make such choices randomly, and such internal ambiguity will increase the amount of random measurement error.

According to social judgment theory, attitudes are not single points on latent attitude continua (Sherif and Hovland 1961; Sherif, Sherif, and Nebergall 1965). Rather, people have what Sherif and colleagues

called "latitudes of acceptance," regions of the attitude dimension they find acceptable, and in which their attitude toward an object falls. Sherif's early work indicated that most people's attitudes toward most objects involved latitudes of acceptance, presumably larger than the single points expressed in the response options of survey questions.

The latitude of acceptance constitutes the region of an attitude scale (or set of response categories) within which a respondent would presumably place himself or herself. The larger the latitude of acceptance, the greater range within which the respondent would potentially respond. In order to report attitudes with large latitudes of acceptance, respondents to survey questions must presumably resolve the ambiguity in such a way as to select a single point on the response dimension. This resolution would seem likely to involve a component of random choice that would not necessarily occur when response categories of survey questions translated directly from accessible attitudinal cues.[5] According to Sherif's view, the most appropriate attitude measurement approach would be to ask respondents to indicate the region of an attitude scale that corresponds to their attitude. Forcing respondents to choose a single point, which is presumably a potentially imprecise point within their latitude of acceptance, may lead respondents to make such choices randomly, even when internal attitudinal cues are relatively unambiguous. Such response ambiguity will increase the amount of random measurement error.

Ambiguity of Response Scale Alternatives

Given the presence of an attitude and some clear internal representation of that attitude, some ambiguities may still remain regarding the expression of the attitude in terms of the response options provided. Random variation in attitude reports is likely to result from ambiguity in the translation of respondent's attitudes to the response continua offered by survey questions. Even assuming unambiguous, accessible internal cues regarding one's attitude, respondents may have difficulty choosing a single point on the response scale provided by the question. Some response errors may, thus, be associated with ambiguous or conflicting judgments regarding the linkage between internal cues and the external response categories.

No matter how clear and unambiguous a respondent's internal attitudinal cues are, he or she must usually express those cues by selecting one of a series of response options offered by the survey question. This entails a process of mapping the internal cues on to the most appropriate response alternative. This mapping process can produce random measurement error in one of two ways. First, the respondent might find that his or her internal cues do not correspond to any of the offered response choices. For example, a respondent who feels that abortion should be legal only if the mother's life is endangered and not under any other circumstances may have difficulty in mapping that view on to the response choices of a question that simply asks whether he or she favors or opposes legalized abortion. When faced with this challenge, respondents may be forced to respond randomly.

Second, the meaning of the response choices may be either clear or ambiguous, and ambiguity would presumably enhance random measurement error. This notion can be illustrated most easily in the context of factual measurement. If a researcher is interested in determining how often respondents smoke cigarettes, they could be asked whether they smoke "constantly, frequently, sometimes, rarely or never." However, the meanings of these response alternatives are clearly ambiguous, particularly in comparison to a question that asks respondents to report the number of cigarettes they smoked in the last 48 hours. In this latter case, the response alternatives are much clearer in their meaning. The more ambiguous the meaning of the response alternatives, the more difficulty respondents should have in mapping their internal attitudinal cues on to those alternatives, and the more random error there will be in responses. In some cases verbal labels may enhance the quality of the responses, but in others they may simply contribute to ambiguity.

HYPOTHESES

These theoretical assumptions regarding sources of random error in attitude measurement can be tested by examining whether variation in question or respondent characteristics is associated with variation

in reliability. In this section we describe a series of question characteristics that may be related to the ambiguity of response alternatives and tendencies for random response by people with no attitudes or ambiguous ones, and which therefore may be related to reliability. We also describe a set of respondent characteristics that may be related to the ambiguity of internal attitudinal cues and abilities to impose clarity on ambiguous stimuli, which are theoretically related to unreliability of attitude responses.

QUESTION CHARACTERISTICS

As indicated above, we assume that attitudes fall along a single, latent continuum, ranging from positive to negative. The reliability of assessments or points on these internal latent continua may be affected by several item characteristics: (a) the number of scale points, (b) whether a midpoint is used, (c) the extent and nature of verbal labeling of response options, and (d) whether a "don't know" response option is explicitly offered.

Number of Scale Points

Offering respondents relatively few response alternatives may not provide enough scale differentiation for reliable mapping of affective reactions toward attitude objects. This raises the question of whether reliability is affected by such imprecision. Consider, for example, a question that offers respondents three response alternatives indicating their favorability toward a government policy: "favor," "neither favor nor oppose," and "oppose." A respondent whose attitude is extremely favorable or unfavorable should readily select one of the extreme alternatives. And a respondent who has neither favorable nor unfavorable feelings would presumably choose the middle alternative. However, a respondent with a relatively weak favorable or unfavorable attitude is confronted with a difficult decision. She or he must choose either the middle alternative, thereby giving the incorrect impression that she or he has no preference or is uncertain, or she or he must choose one of the extreme alternatives, giving the impression that she or he has stronger feelings than is in fact the case.

Choices made by such respondents when confronted with too few response categories may very likely be random. Such respondents would probably prefer to have available response alternatives indicating weak, moderate, and strong positive and negative evaluations, in part because these are categories that people often use to describe attitudes and opinions. If such additional response options were offered, the random guessing, or what Lehmann and Hulbert (1972) called "rounding error," typical of responses to few response categories, would presumably be reduced. This reasoning supports the claim that scales with more response alternatives will be more reliable than those with fewer.[6]

There is probably a limit to the benefit of adding response categories or scale points. Because people probably differentiate between weak, moderate and strong feelings toward attitude objects, 7-point response scales seem preferable to shorter ones, and among shorter ones, more points would seem likely to be associated with greater reliability. However, once scales grow much beyond seven points, the meaning of the specific scale points presumably becomes increasingly ambiguous (Miller 1956). And, as we argued above, ambiguity in the meanings of scale points is likely to increase random measurement errors. Therefore, the relation between the number of scale points and reliability may be curvilinear. That is, reliability may increase up to 7-point scales (and possibly somewhat beyond), and may level off or decrease thereafter, so scales with 10 or more points may be no more reliable than 7-point scales.

It is often stated that the reliability of scales increases with the number of scale points used (e.g., Jahoda, Deutsch, and Cook 1951). And although there is considerable opinion in favor of this principle (e.g., Symonds 1924; Champney and Marshall 1939; Ferguson 1941; Murphy and Likert 1938), the evidence in support of it is virtually nonexistent. Bendig (1953) found that interrater reliabilities (computed by intraclass methods) indicated equal reliability for rating scales having 3, 5, 7, or 9 categories, but a decrease in reliability for 11 categories. Komorita (1963) analyzed internal consistency reliabilities for 14-item composite indexes that used 2-point versus 6-point scales, reporting no difference in the composite reliabilities. Matell and Jacoby (1971) found that reliability and criterion-validity are

independent of the number of scale points used for Likert-type item, arguing that regardless of the number of response categories employed, "conversion to dichotomous or trichotomous measures does not result in any significant decrement in reliability or validity." Komorita and Graham (1965) found that for relatively homogeneous items, composite reliability does not increase with the number of scale-points, but among sets of heterogeneous items, a gain in composite reliability can be obtained by using 6 versus 2 scale-points for the items. Using simulated data, Lissitz and Green (1975) reproduce the Komorita and Graham results, but suggest that 7 scale points may not be optimal. They find that after 5 scale points there is a leveling off of the increase in reliability. Similar results were obtained by Jenkins and Taber (1977).

All of the above-cited research has focused on the effects of number of response categories on the reliability of linear composites or on interrater reliability in nonsurvey measurement settings. And, although this type of research is of interest (see review by Cox 1980), it is less pertinent to the question of the reliability of single survey items. At the present time there is considerably more interest in the behavior of single survey questions under different conditions of measurement, and the study of items individually may be much more relevant to discussions of response errors in surveys. Of course, most of the above-cited research was published prior to the development of routine methods for estimating the reliability of single items, and hence little information exists on this question.

Some survey research evidence exists on this topic. Andrews and Withey (1976), for example, compared 7- and 3-category response options in the measurement of well-being, noting that "seven-point scales provide more sensitive indications of respondent's feelings" compared to 3-point scales. They conclude that 3-category scales "capture only 80-90 percent of the total variation, whereas seven-category scales capture virtually 100 percent of it."

Andrews (1984) compared response scales of 2, 3, 4 and 5, 7, 9 to 19 and 20+ categories in terms of estimated reliability, validity and method variance. He concluded that the number of response categories had larger effects on "data quality" than other aspects of question

design. Specifically, he concluded that "as the number of answer categories goes up, data quality goes up, i.e., validity tends to increase and residual error tends to decrease." Although he found reliability generally increased with more response categories, 3-point scales were found to be less reliable than 2- or 4-to-5-category response scales. Also, he reports that there is no clear tendency for method variance to increase with the use of more response categories. Unfortunately, these results are less informative than desirable, since Andrews analyzed a pool of survey questions measuring a wide range of content, including subjective variables as well as reports of factual information. Several of the questions in his set of results involved reports of frequencies of behavior, for which the number of days per month was requested, including roughly 30 response alternatives. Thus these results may not bear on the question of interest here.

In an experimental study carried out in the 1984 General Social Survey on the measurement of confidence in social institutions, Smith and Peterson (1985) reported that when compared to 3-point scales, 7-point response scales do not produce higher interitem correlations. In fact, they argue that 7-point scales produce greater amounts of respondent error than do their 3-point counterparts. Smith and Peterson do not, however, estimate the reliabilities of their two types of scales.

Scales with Midpoints

Among response scales involving fewer than seven points, it is possible to take the additional step in differentiating those that have a midpoint (odd numbers of scale points) from those that do not (an even number). Some respondents have no attitude toward an object or have genuinely ambivalent feelings that are equally positive and negative. These respondents would presumably prefer to place themselves at the middle of the evaluative continuum. However, if they are faced with a response scale with an even number of response alternatives, there is no middle alternative that would accurately reflect their lack of preference. These respondents would therefore be forced to choose between representing themselves as weakly favorable or weakly unfavorable toward the attitude object. This choice may often be random,

so offering a middle alternative to respondents would presumably increase the reliability of responses. Therefore, scales with odd numbers of response categories may be more reliable than scales with even numbers of response alternatives.

Moreover, instead of responding randomly, respondents who have no attitudes may be just as likely to adopt a *satisficing* strategy, responding in a safe, nondescript way. In other words, such persons may seek a satisfactory response to the question rather than an optimal one, a tendency referred to as satisficing (Simon 1977). If this is the case, the midpoint might be chosen more often, and more reliably so, among such persons, and one would expect that if a middle alternative is provided, such questions may show higher reliabilities.

As noted above, Andrews (1984) found 3-point scales to be less reliable than 2-point and 4-to-5-point scales, which is contrary to this hypothesis. Also, contrary to our predictions, Andrews found that scales with midpoints were not more reliable than those scales without them. Again, however, because Andrews analyzed attitude and fact measures simultaneously, it is difficult to reach any conclusion from his results about the impact of item characteristics on the reliability of attitude measurement.

Verbal Labeling

Response scales composed of numbers probably involve some inherent ambiguity of meaning. Attaching verbal labels to numbered response options, or defining them in some other way, probably clarifies the meanings of these alternatives. There is probably some ambiguity inherent even in verbal definitions, so verbal labeling would not be expected to completely eliminate unreliability due to response scale ambiguity. However, attaching labels may reduce random reporting somewhat.

Several studies provide support for this hypothesis. Bendig (1953), Madden and Bourdon (1964), Finn (1972), Peters and McCormick (1966), and Zaller (1988) all found that increasing the proportion of the scale points that are verbally labeled increased item reliability. Surprisingly though, Andrews (1984) found reliability to be lower for fully labeled scales than for partially labeled ones.

Offering "Don't Know" Response Options

Converse's (1964) original description of the factors that produce nonattitude random reporting emphasized the interpersonal dynamics that operate during face-to-face interviews. According to his thesis, respondents feel some pressure to express opinions because they assume that interviewers want them to do so. Making it clear to respondents that interviewers do not expect them to offer opinions by providing explicit "don't know" response options, or by introducing "no opinion" filters, may therefore reduce the number of nonattitudes expressed.[7]

Evidence for this hypothesis exists in several forms. In some research, respondents were asked whether they had an opinion on a given topic before they were asked what their opinion was. In other research the question included the "don't know" option along with the other response categories. Both types of studies have found that legitimizing nonattitudes with these procedures increased the numbers of people who said they "don't know" or have "no opinion" (Schuman and Presser 1981; Bishop, Oldendick, and Tuchfarber 1983). The only study of which we are aware that tested the effects of offering "no opinion" options on reliability was the study by Andrews (1984). Consistent with our predictions, offering respondents the option to say "don't know" increased the reliability of attitude reports.

Interpreting this result is not straightforward. Respondents may choose a "don't know" alternative because they truly have no attitude or for other reasons. As noted above, they may lack the motivation to probe their thoughts and arrive at a reasoned response, or they may wish to keep their opinions to themselves, and may select this choice as a way of avoiding the question. Or, they may be aware of the region of their attitude in terms of the response scale and may be genuinely uncertain of exactly which point represents it best. For the latter type of respondent, the "don't know" response represents uncertainty of the mapping of their attitude to the response scale (Coombs and Coombs 1976). Therefore "don't know" response options may increase reliability by filtering out respondents with wide latitudes of acceptance/rejection.

RESPONDENT CHARACTERISTICS

A wide array of respondent characteristics may affect the randomness of response. Perhaps the most important of these are respondent motivation and cognitive ability. However, such respondent characteristics are difficult to measure in the context of survey interviews, and we must therefore rely on more easily assessed proxies for such factors. We consider two such variables: level of schooling and age.

Schooling

Access to greater amounts of schooling in modern society requires greater cognitive abilities and prior verbal learning. Accordingly, school attendance presumably promotes these same traits and provides considerable practice with multiple choice questions. Schooling encourages persons to think about social and political affairs, and thus, it reduces the existence of nonattitudes and the random reporting that results from them. At the same time, schooling may be thought to reduce the ambiguity of internal cues by providing experiences that help persons recognize the nature of their own feelings with respect to these issues. On the other hand, because schooling increases intellectual flexibility and the likelihood that persons will learn to think for themselves, this interpretation is not necessarily straightforward. For some persons additional schooling may in fact promote a higher tolerance for ambiguity and an overall reluctance to quickly form an opinion on a complex social issue. Finally, given the experiences that schooling provides in the way of responding to multiple choice examinations, more schooling undoubtedly reduces random responding. Presumably such learning assists "schooled" respondents in readily translating their attitudes into the categories of the response scales used in surveys, and therefore less randomness results from this source. Therefore, the factors that lead to less reliable survey response —the existence of nonattitudes, ambiguity in internal cues, and ambiguity of external cues—all seem to be correlated negatively with amount of schooling.

Previous research uniformly reports less reliability of measurement in less educated respondents (see e.g., Converse 1974, 1980). Judd and

Milburn (1980) and Judd, Krosnick, and Milburn (1981) found greater disturbance variance for respondents with less schooling, but these estimates confounded unreliability with attitude-specific variance.

Age

Advancing age may lead to less measurement reliability because of mental decay, decreased judgment, and poorer memories. However, previous research on this issue provides conflicting results. Kogan (1961) found weaker correlations between similarly phrased attitude items among older than younger age groups, suggesting the possibility of greater random error, and thus greater attenuation in bivariate relationships. Andrews and Herzog (1986) found true-score variance tended to decline with age, whereas method variance and random error variance increase. This would suggest that reliability will decline with increasing age. Interestingly, Andrews and Herzog's (1986) results were not linear with age, declining systematically at about age 55, remaining relatively stable thereafter. However, other evidence (e.g., Rodgers and Herzog, 1987a, 1987b) suggested that measurement errors were no greater for older respondents than for younger ones.

DATA AND METHODS

THE SURVEYS

Panel data sets were selected for use in this study if they were national in scope, if they had a minimum of roughly 200 respondents, if they had at least three waves of data, and if they had a sufficient number of attitude questions to make their analysis worthwhile. We found five extant panel data sets that met these criteria: (a) the 1956, 1958, and 1960 National Election Study (NES) Panel (n = 1,132), (b) the 1972, 1974, and 1976 NES Panel (n = 1,320), (c) the 1980 NES Panel (n = 769), (d) the 1973 reinterview subsample of the General Social Survey (n = 195), and (e) the 1974 reinterview subsample of the General Social Survey (n = 195).

National Election Panel Studies

Every two years since 1952 (except 1954), the University of Michigan's Institute for Social Research has interviewed a representative cross-section of Americans to track national political participation. On the years of presidential elections, a sample is interviewed before the election and is reinterviewed immediately afterward. In the non-presidential election years only postelection surveys are conducted. Data are obtained from face-to-face interviews with national full-probability samples of all citizens of voting age in the continental United States, exclusive of military reservations, using the Survey Research Center's multistage area sample (see Miller, Miller, and Schneider 1980).[8] The sample sizes typically range between 1,500 and 2,000.

Of the respondents interviewed in 1956, 1,132 of them were reinterviewed in 1958 and again in 1960. The 1958 and 1960 panel questionnaires were the same as those used in the 1958 and 1960 cross-sections respectively. This design afforded only a small number of items that were replicated in all three studies. Of the respondents interviewed in 1972, 1,320 were successfully reinterviewed in 1974 and again in 1976. Again, the questionnaires for these reinterview surveys were the same as those used for the cross-sectional samples interviewed at those times. The data from the 1970's panel design, however, yielded many more replicate attitude questions. In the 1980 National Election Panel Study, 769 respondents were reinterviewed at roughly 4-month intervals, beginning in January and ending in November (see Markus 1982).[9]

General Social Survey Reinterview Studies

The General Social Survey (GSS) is an annual cross-sectional survey of the noninstitutionalized residential population of the continental United States aged 18 and over (National Opinion Research Center 1987). It has been conducted nearly every year since 1972 on approximately 1,500 respondents per year. The purpose of the GSS has been to monitor social trends in attitudes and behavior. The GSS does not ordinarily include a panel component, however, in 1972,

1973, 1974, 1978, and 1987 such a design was included. In the 1973 and 1974 reinterview studies, three waves were involved, and we used these data here (see Smith and Stephenson 1979). In the 1973 study, the GSS attempted to reinterview a random subset of 315 respondents to the initial survey, of which 227 completed a second interview and 195 completed a third. We analyzed the data from the 195 cases surviving the three waves of the study, or 62% of the original target sample. In the 1974 study, attempts were made to reinterview 291 of the original GSS respondents, of which 210 were reinterviewed a second time, and 195 a third. Again, we analyzed the data from the 195 cases with complete data for all three waves, or 67% of the target sample. The average interval between the first and second waves in the 1973 study was 46.9 days, the average interval between the first and third waves was 80.2 days. In the 1974 study the average intervals between first and second was 46.4 days and between the first and third, 78.9 days. The 1973 reinterview study included 44 questions that were common across all three waves, 14 of which were attitude questions. The 1974 study included 19 questions in the second and third waves, common to the initial survey, 4 of which were attitude questions. The initial GSS interviews were conducted face-to-face, and reinterviews were by telephone (see Smith and Stephenson 1979).

MEASURES

We analyzed 96 measures (m) of attitudes from these five reinterview surveys: (a) 1956 to 1960 NES, $m = 9$; (b) 1972 to 1976 NES, $m = 51$; (c) 1980 NES, $m = 23$, and (d) the combined 1973 and 1974 GSS surveys, $m = 13$.[10] We restricted our analysis to attitude measures, excluding measures of perceptions, beliefs, self-evaluations, and factual material (but see Alwin 1989). The response options for these measures ranged from agree-disagree type questions, which vary in number of response options, the extent of labeling, and so on, to rating scales involving any number of scale points. The longest scales are the "feeling thermometers" which have 9 scale points.[11] There is a fair number of 7-point scales, and several others involving 2, 3, 4, or 5 categories. All response scales label the extreme categories with verbal

anchors, but scales vary in the extent to which they label more than the extreme categories. Finally, there are a few forced-choice questions, which ask the respondent to choose between two or more statements in terms of their reflection of his or her attitude. This form of question occurs primarily in the GSS panels.

ANALYSIS

We employ a class of just-identified simplex models that specify two structural equations for a set of 3 over-time measures of a given variable y_t:

$$y_t = \tau_t + \varepsilon_t$$
$$\tau_t = \beta_{t\tau t-1}\, \tau_{t-1} + \upsilon_t$$

The first equation represents a set of measurement assumptions, indicating that the over-time measures are assumed to be tau-equivalent except for true attitude change and that measurement error is random (see Alwin, 1988). The second equation specifies the causal processes involved in attitude change over time. This model assumes that the system is in dynamic equilibrium and that this equilibrium can be described by a lag – 1 or Markovian process in which the distribution of the true variable at time *t* is dependent only on the distribution at time *t* – 1 and not directly dependent on distributions of the true variable at earlier times.

Using this model we obtained estimates of the proportion of response variance that can be attributed to "true" attitudes, that is, σ_τ^2/σ_y^2. These were obtained by estimating the parameters of structural equation models for three-wave panel data (see Heise 1969; Jöreskog 1970, 1974; Werts, Jöreskog, and Linn 1971; Wiley and Wiley 1970).[12] All reliability estimates were obtained using Jöreskog and Sörbom's (1986) LISREL computer program.[13] These reliability estimates were then used as input to a secondary analysis of the influences of question and respondent characteristics on attitude reporting reliability. For all measures we estimated reliability using two different sets of assumptions. First, we estimated these models assuming that the reliability was constant over occasions of measurement (see Heise 1969). Second, we estimated these models relaxing this assumption, allowing

reliability to differ from occasion to occasion (see Wiley and Wiley 1970). We found that in general very few differences existed in the separate reliability estimates of the second model, and because the reliability of the second time point equals the single reliability estimated in the Heise model, we decided to use this single estimate of reliability (see Alwin 1989).[14]

The analysis we present here does not examine the interaction between the characteristics of questions and respondents. Although this is of some theoretical interest, several factors support our decision to exclude such analytic goals from the present analysis. First, because of the quasi-experimental nature of our design, such an analysis would involve considerable loss of power, both because of reduced sample sizes for examining effects of question differences on reliability within subgroups of respondents and because of reduced numbers of questions within a given survey. Second, previous research has found little evidence of interaction between respondent characteristics and methodological features of survey questions. Schuman and Presser (1981), for example, found little evidence that education interacts with question characteristics in affecting response distributions. And more recently, investigators (Rodgers, Andrews, and Herzog 1989) reported an analysis of more than 100 survey measures, concluding that few differences in data quality existed for subgroups of respondents defined by age, education and a variety of other characteristics. Thus we expect that by ignoring the possibility of interactions of question and respondent characteristics in their influence on the reliability of attitude reports, our present analysis does not oversimplify too greatly any possible interactions that might exist.

RESULTS

Our main results consist of average levels of reliability presented by categories of the question and respondent characteristics of interest. Question characteristics are confounded with one another and with certain aspects of the panel designs. In order to examine the influences of design and item characteristics on item reliability, we employed Multiple Classification Analysis (MCA), a method of multiple regres-

TABLE 1: Design Characteristics of the Five Panel Data Sets Included in the Present Study

Data Set	*House*	*Reinterview Period*	*Number of Questions*	*Sample Size*	*Average Reliability*
1950s NES	SRC[a]	24 months	9	1,132	.505
1970s NES	SRC	24 months	51	1,320	.505
1980s NES	SRC	4 months	23	769	.692
1973-74 GSS	NORC[b]	2 months	13	195	.721
F ratio		16.54			
Degrees of freedom		(3, 92)			
p value		$p < .001$			

a. SRC = the Survey Research Center, University of Michigan.
b. NORC = the National Opinion Research Center.

sion using categorical predictors that assesses the effects of categorical variables on continuous dependent variables, and controls for the influences of other variables (Andrews, Morgan, Sonquist, and Klem 1973).[15] This approach assesses the nonlinear, additive effects of predictor variables.

DESIGN CHARACTERISTICS

Table 1 presents the design characteristics of the 5 panel studies used here. The table also presents the average reliability for each of the five studies, indicating statistically significant differences over studies ($F[3, 92] = 16.5\, p < .001$).[16] As these results indicate, estimated reliability is generally highest for those studies with the shortest time period between interviews. This is presumably due, at least in part, to a set of factors referred to by Moser and Kalton (1972, p. 353), that "respondents may remember their first answers and give consistent retest answers, an action which would make the test appear more reliable than is truly the case." Over longer time periods reliability is lower and presumably more accurately estimated. In the following analyses of the effects of question characteristics on estimated reliability, we control for the time period between reinterviews, in order to remove these influences.

TOPIC OF THE QUESTION

In addition to differences between the five studies in levels of estimated reliability directly attributable to the length of the time interval between interviews, they also differ in reliability due to the topics covered in that particular survey. These attitude measures addressed six types of content: (a) social and political attitudes on specific social issues, including federally guaranteed employment, protecting the rights of people accused of committing crimes, government policies involving racial minorities, the role of women in society, the conditions under which women should have a right to an abortion, the dissemination of birth control information, civil liberties and government spending (m = 33); (b) political efficacy and alienation (m = 16); (c) the evaluation of social groups (m = 17); (d) the evaluation of political candidates (m = 16); (e) party identification (m = 9); and (f) political ideological liberal-conservatism (m = 5).[17]

Table 2 shows that the topic categories employed here are not uniformly distributed across the five studies. Virtually all of the GSS questions involve policy content (12 of 13). Thus all of the remaining categories are drawn from the Election Studies.[18] Moreover, virtually all of the "political efficacy," "social groups," and "ideology" questions are from the 1970's NES study.

Table 2 presents the reliability estimates by topic category within each of the five data sets. As shown, in some cases there are insufficient numbers of questions in a particular topic category to draw any conclusions about the influences of topic on reliability within any particular study. For the 1970s and 1980s NES panels, there are some important differences between topic areas within each of these data sets (1970s: $F[5, 45] = 5.85$, $p < .001$; 1980s: $F[2, 18] = 6.05$, $p = .01$).[19] In addition, the table presents the estimated reliabilities by topic, using the data combined over the five studies. These results, adjusted for differences between studies in the time interval between reinterviews, collectively show meaningful differences by topic ($F[7, 88] = 10.37$, $p < .001$).

The results regarding topic suggest that some content domains can be more reliably measured than others. Candidate ratings, ideological assessments, and measures of party identification are the most reliable,

TABLE 2: Estimates of Question Reliability by Data Set and Six Categories of Question Topic

	1950s NES	*1970s NES*	*1973/1974 GSS*	*1980s NES*	*Total*[a]
Policy	.458	.490	.725	.613	.543
	(8)	(6)	(12)	(7)	(33)
Efficacy	—	.424	—	—	.477
		(16)			(16)
Groups	—	.481	.663	.395	.521
		(15)	(1)	(1)	(17)
Party	.885	.594	—	.788	.714
	(1)	(5)		(3)	(9)
Candidate	—	.701	—	.747	.724
		(5)		(11)	(16)
Ideology	—	.588	—	.654	.636
		(4)		(1)	(5)
Total	.505	.505	.721	.692	.579
	(9)	(51)	(13)	(23)	(96)
F ratio		5.85		6.05	10.37
Degrees of freedom		(5, 45)		(2, 18)	(7, 88)
p value		$p < .001$		$p = .01$	$p < .001$

a. Adjusted for time between waves

and social policy attitudes, political efficacy, and attitudes toward social groups are measured significantly less reliably. We hypothesized that the differences between types of attitudinal content in reliability of measurement were due to the fact that different forms of survey questions were used for these different kinds of content, and that these question forms may be inherently different from one another, as argued above, in the magnitude of random error they engender.

The most obvious differences are those involving the number of scale points. Party identification, for example, was measured using an unfolding format in which respondents were first asked whether they considered themselves to be Republicans, Democrats, or Independents (see Alwin and Krosnick 1989). People reporting an identification with one of the two parties were then asked whether they did so strongly or weakly. People who said they were Independent were asked whether they leaned toward one party or another. As a result,

respondents were segmented into seven groups along a continuum ranging from strong Republican to strong Democrat. This unfolding approach presumably makes it very easy for respondents to understand the meaning conveyed by the responses they provide to each question, so they end up being highly reliable. Similarly, the "liberal-conservative" ratings were acquired by asking respondents to place themselves on a fully labeled 7-point scale (see Alwin and Krosnick 1989). In contrast to this, the policy questions were most often 7-point scales, with only the end-points labeled with words. The format may involve more ambiguity in meaning of the mid-range, and therefore may increase random error. Similarly, many of the efficacy or alienation questions are dichotomous or trichotomous and are likely to be less reliable than continuous response scales because random error in reports of attitudes near the midpoint of the attitude continuum may cause these responses to oscillate from one side to the other. If these respondents were given the opportunity to express slight leanings in one direction or the other, as we argued above, their reports would potentially be more reliable.

Because of these expectations regarding the confounding of attitude content with attitude questions, we find it necessary to control for question content to the extent possible in assessing the effects of question characteristics. Also, as indicated, because of the relation of the length of the reinterview period to reliability, we also control for these differences, either by selection or by statistical adjustment.

QUESTION CHARACTERISTICS AND RELIABILITY

In this section we present results relating estimates of item reliability to various question characteristics: (a) number of response categories, (b) presence of a middle alternative, (c) the extent of labeling of response options, and (d) the explicit offering of a "don't know" alternative.

NUMBER OF RESPONSE CATEGORIES

It was hypothesized that reliability of attitude reporting will improve as the number of response options increases, up to a point, and

decrease beyond that. We noted that the existing research literature on this issue indicates that 7- to 10-point scales may be the most reliable. In the present analysis we exclude the GSS questions because virtually all involved only two response categories. Within the NES panels, most of the variation in number of response categories occurs among the rating scales (m = 63), although there is some variation (2- vs. 5-category scales) among the agree-disagree questions (m = 14).

Table 3 presents the results of a comparison of the estimated reliability of questions involving differing numbers of scale points for "agree-disagree" and "rating scale" questions in the NES panels, controlling for the time-interval of reinterview and the topic of the question.[20] In this analysis we exclude the measures of party identification resulting from use of the unfolding format in that they are not true rating scales, that is, they result in 7-point scales, but they are not 7-point rating scales (see Alwin and Krosnick 1989).

The results for the number of scale-points confirms some of our predictions, but the results are mixed. On the one hand, among the agree-disagree questions there are no statistically significant differences between the 2- and 5-category response scales.[21] On the other hand, among the rating scales, from 3-point response scales upward there is a generally monotonic increase in reliability, with no perceptible differences between the 7- and 9-point response scales. The analysis combining the agree-disagree and rating scale formats, which adjusts for differences in the reinterview period and topic of the question, shows, with some notable exceptions, that there is a general monotonic increase in reliability with greater numbers of response categories. The 2-category scales are a major exception to this pattern, as they have relatively reliable responses. We suspect this is because 2-category questions unambiguously measure the direction of attitudes only, with no pretense of measuring intensity, whereas 4 and more category response scales presumably are intended to measure both direction and intensity. The direction of attitude responses may in fact be more reliably assessed than the intensity of responses (see Alwin 1991). For reasons given in the next section, 3-category response scales were found to be less reliable, consistent with Andrews's (1984) finding. Also, with the exception of the 5-point rating scales (see below) there is a clear pattern of increasing reliability with more

TABLE 3: Estimates of Question Reliability in NES Panels by the Number of Response Categories by Response Scale Types, Adjusting for Design Characteristics

	Agree-Disagree		*Rating Scales*			*Total*		
	n	*Unadjusted Mean*	n	*Unadjusted Mean*	*Adjusted Mean*	n	*Unadjusted Mean*	*Adjusted Mean*
Two	6	.480	–	–	–	7	.458*	.541
Three	–	–	20	.410**	.466**	22	.427**	.477**
Four	–	–	4	.528	.561	4	.528	.508
Five	8	.458	–	–	–	8	.458*	.492
Seven	–	–	10	.588	.619	10	.588	.572
Nine	–	–	29	.669**	.615*	29	.669**	.610**
Total	14	.467	63	.565	.565	80	.546	.546
F ratio		0.75		21.46	6.03		14.88	3.47
Degrees of freedom		(1, 12)		(3, 59)	(7, 55)		(5, 74)	(9, 70)
p value		$p = .40$		$p < .001$	$p = .001$		$p < .001$	$p = .007$

*$p < .05$; **$p < .01$

response categories in the combined agree-disagree and rating form results.

MIDDLE ALTERNATIVES

We also hypothesized that questions with an odd number of scale categories—those with a middle alternative—would show greater reliability than those with an even number. Thus we predicted a "saw-toothed" pattern of reliabilities, with odd-numbers of scale points showing higher reliabilities. But our results show the opposite pattern. Three-point scales are less reliable when compared to the 2- and 4-point scales. It is also worth noting that the 5-point agree-disagree are no more reliable than the 4-point rating scales. Middle alternatives may in fact lower reliability of measurement. Middle alternatives may become more valuable in longer response forms, such as 7-point rating scales, where they can serve as an anchor for opinion (see Saris, 1988).

VERBAL LABELING OF RESPONSE OPTIONS

We hypothesized that the more verbal labeling is used for the response categories of survey questions, the greater will be the estimated reliability of measurement. We argued that labels reduce ambiguity in translating attitudes into the categories of response scales. In the present data sets, labeling of response categories was extensive. The only case in which variation existed in the labeling of response options was within the 7-point scale questions, as used in the National Election Studies. All these response scales label the end-points, so this amounts to a test of the linkage between reliability and the extensiveness of labeling. That is, we compared the reliability of scales that label the end-points only to those that provide complete labeling of response categories.

Table 4 presents a comparison of reliabilities of fully labeled 7-point scales with those in which only the end-points are labeled. These results indicate significant differences in reliability in favor of fully labeled response scales ($F[2, 10] = 9.39, p < .05$), confirming our expectation that more labeled categories produce higher reliabilities.[22]

TABLE 4: Estimates of Question Reliability by the Method of Labeling Response Categories Among 7-Point Scales, Adjusting for Design Characteristics

	n	*Unadjusted Mean*	*Adjusted Mean*
Fully labeled	5	.783**	.783*
Only endpoints labeled	8	.570**	.570*
Total	13	.652	.652
F ratio		10.04	9.39
Degrees of freedom		(1, 11)	(2, 10)
p value		$p < .01$	$p < .05$

*$p < .05$; **$p < .01$

These results provide support for the practice of labeling response scales extensively.

EXPLICIT OFFERING OF "DON'T KNOW" OPTION

We predicted that by offering an explicit "don't know" option, the nonattitude problem would be directly confronted, and random responding would be reduced. In the data set assembled here, it is difficult to obtain an independent assessment of this hypothesis. There is variation in this characteristic within agree-disagree questions, in that 8 such questions include an explicit "don't know" option and 6 do not. However, as we pointed out earlier with respect to Table 3, this comparison confounds topic and number of scale points with whether a "don't know" category is provided.[23]

There is some variation among the thirteen 7-point rating scales in whether a "don't know" response option was explicitly presented —10 offer a "don't know" option and 3 do not. This comparison is given in Table 5. These results show that, contrary to our hypothesis, the explicit offering of a "don't know" option does not appear to produce an improvement in reliability. Not only do these results provide no support for our hypothesis, the results are in the opposite direction. Furnishing a "don't know" option appears to lower the reliability, a result not expected and not consistent with previous research (Andrews 1984). However, given the limited amount of

TABLE 5: Estimates of Question Reliability by the Method of Obtaining Don't Know Responses, Adjusting for Design Characteristics

	Agree/Disagree			*7-Point Ratings*	
	n	*Unadjusted Mean*	*Adjusted Mean*	n	*Unadjusted Mean*	*Adjusted Mean*
DK offered	8	.458	–	10	.588**	.606
DK not offered	6	.480	–	3	.863**	.805
Total	14	.467	–	13	.652	.652
F ratio		0.75			15.95	4.51
Degrees of freedom		(1, 12)			(1, 11)	(2, 10)
p value		$p = .40$			$p < .01$	$p = .06$

*$p < .05$; **$p < .01$

information available in the data sets used here, it is risky to draw a firm conclusion from these results.

Perhaps people who are attracted to the "don't know" filter when it is offered would have placed themselves extremely reliably at the scale midpoint had the "don't know" filter not been offered. That is, these people might be highly reliable, so removing them (by offering them a "don't know" filter) lowers the average reliability.

RESPONDENT CHARACTERISTICS

We present reliability estimates for categories of schooling based on all of the GSS and NES studies. Because of the similarity of design, we group the 1950s and 1970s panels, as well as the 1973 and 1974 GSS reinterview data sets. In the analysis of reliability by age, we rely on the NES data only.[24] Tables 6 and 7 present these results for 4 categories of schooling and 7 age categories respectively.

RELIABILITY AND SCHOOLING

In order to analyze differences in reliability by level of schooling, we partitioned each data set into four categories of schooling: (a) those with less than completed secondary schooling, (b) those with com-

pleted secondary schooling, (c) those with more than completed secondary schooling but with no college degree, and (d) those with a college degree or more. We then analyzed variation in attitude reliabilities over these categories. These results are given in Table 6.[25]

These results show, as expected, a systematic increase in levels of reported attitude measurement reliability. This supports our previously hypothesized contention that schooling provides experiences that reduce the tendency to report attitudes randomly. For a variety of reasons schooling reduces the amount of unreliability of attitude measurement. This finding fits nicely with the interpretation made by Converse (1964) more than 20 years ago with respect to the differences in the responses of *elite* and *mass publics* to survey attitude questions. And this provides further confirming evidence for the interpretation of greater randomness in the responses of mass publics (see Judd and Milburn 1980; and Judd, Krosnick, and Milburn 1981). Although the results for schooling in the GSS data are not significant, the pattern of coefficients is consistent with the hypothesis and the differences observed in the NES panels.

RELIABILITY AND AGE

In examining the relation of reliability to age, we analyzed the data separately by categories of "length of the reinterview period," in that we grouped the 1950s and 1970s results. We partitioned each NES data set into seven age categories: (a) 18 to 25, (b) 26 to 33, (c) 34 to 41, (d) 42 to 49, (e) 50 to 57, (f) 58 to 65, and (g) 66 and above, and analyzed variation in reliabilities for all available attitude measures across groups. These results are given in Table 7.

These results show no overall statistically significant differences in attitude reporting reliability by age. In the 1980s NES panel, there is a significant decline in reliability in old age, consistent with one hypothesis advanced in the literature. And, although such a pattern is perhaps evident in the combined 1950s and 1970s data, the overall differences are small, and as indicated, not statistically significant. Still, the overall weight of the evidence suggests a nonmonotonic relation to age, with the oldest age category showing a lower reliability. In both of the NES remeasurement designs there is a systematic

TABLE 6: The Relationship Between Schooling and Estimates of Reliability of Measurement for Attitude Questions, National Election Study Panels, 1956 to 1960, 1972 to 1976, 1980, and General Social Survey Reinterview Panels 1973, 1974

Level of Schooling	*Sample Size*		*Number of Items*	*Mean*	*Deviation From Grand Mean*
NES 1950s, 1970s[a]	1950s	1970s			
0-11 years	365	301	59	.462	−.045*
12 years	266	343	59	.494	−.013
13-15 years	101	204	59	.531	.025
16+ years	80	193	58	.540	.034
Total	812	1,041	235	.507	.000
F ratio			2.48		
Degrees of freedom			(3, 231)		
p value			$p = .06$		
η		.177			
NES 1980s[b]					
0-11 years	117		22	.609	−.089**
12 years	212		23	.657	−.040
13-15 years	119		23	.753	.055
16+ years	122		23	.767	.069*
Total	570		91	.697	.000
F ratio			4.19		
Degrees of freedom			(3, 87)		
p value			$p < .01$		
η		.355			
GSS 1973, 1974[c]	1973	1974			
0-11 years	57	45	11	.686	−.020
12 years	63	56	11	.710	.004
13-20 years	54	73	11	.722	.016
Total	174	174	33	.706	.000
F ratio			0.15		
Degrees of freedom			(2, 30)		
p value			$p = .86$		
η		.100			

a. For the NES 1950-1970 analysis ECON PCY was excluded from Education 16+ years.
b. For the NES 1980s analysis Baker was excluded from Education 0-11 years.
c. For the GSS 1973-1974 analysis ABPOOR, SPKSOC were excluded.
*$p < .05$; **$p < .01$

TABLE 7: The Relationship Between Age and Estimates of Reliability of Measurement for Attitude Questions, National Election Study Panels, 1956 to 1960, 1972 to 1976, and 1980

Level of Schooling	*Sample Size*		*Number of Items*	*Mean*	*Deviation From Grand Mean*
NES 1950s, 1970s[a]	1950s	1970s			
18-25	62	173	60	.511	-.009
26-33	169	194	60	.530	.010
34-41	179	153	60	.507	-.013
42-49	139	163	60	.530	.010
50-57	100	130	60	.542	.022
58-65	71	106	60	.526	.006
66-83	67	110	59	.492	-.027
Total	787	1,029	419	.520	.000
F ratio			0.37		
Degrees of freedom			(6, 412)		
p value			$p = .90$		
η		.073			
NES 1980s[b]					
18-25	97		23	.725	.033
26-33	125		23	.744	.052
34-41	92		23	.706	.014
42-49	54		23	.728	.036
50-57	72		23	.708	.015
58-65	65		23	.634	-.059
66-83	59		21	.593	-.100*
Total	564		159	.692	.000
F ratio			1.39		
Degrees of freedom			(6, 152)		
p value			$p = .22$		
η		.228			

a. For the NES 1950s, 1970s analysis Women was excluded from age 66-83.
b. For the NES 1980s analysis Bush and McGovern were excluded from age 66-83.
*$p < .05$; **$p < .01$

decline in reporting reliability from the age 50 to 57 group, with the 66 to 83 group showing itself to be significantly different ($p < .05$) from the grand mean in the 1980 NES data.

Alwin and Krosnick (forthcoming) show that such a relationship between reliability and age can lead to erroneous conclusions regarding age differences in bivariate relationships among variables (also see Krosnick and Alwin 1989). They show that even this slight decline in reliability in old age can make older persons appear to be relatively less persistent in their attitudes than somewhat younger age groups, where persistence is gauged in terms of relationships among variables measured over time in panel studies (see Sears 1981). These findings underscore the conclusion that measurement reliability differences between populations and/or subpopulations can lead to potentially erroneous conclusions if not taken into account in the analysis.

DISCUSSION AND CONCLUSION

To the extent that characteristics of survey questions are linked to the estimated reporting reliability of attitude measurement, there may be some empirical basis for instructing survey researchers in the development of questions for use in survey interviews. And to the extent that the characteristics of respondents may be linked to levels of reliability, survey researchers can be informed about which subpopulations need special attention in the reduction of errors.

We conclude from this analysis that the reliability of attitudinal survey measures is affected to some extent by the design and format of survey questions and to some extent by the characteristics of respondents themselves. Some of our expectations regarding the influence of survey question characteristics on reporting reliability were confirmed, but others were not. We found, as expected, that response scales with more categories are the most reliable. Among 7-point scales, those that are fully labeled were found to be more reliable than those not so labeled. We found, contrary to expectations based on previous theory and research, that reliability does not seem to be enhanced by explicitly offering a "don't know" option.

One major difference in reliability reported here involves the question content. We find that the measurement of sociopolitical orientations that are more ideological in content, for example, "ideological" self-placements, party identification and candidate preferences are

estimated to be the most reliable, whereas those measures assessing attitudes toward policy issues, those that assess attitudes toward social groups, and those measures seeking expressions of political efficacy or alienation are the least reliably reported. These results might be due to differences in question characteristics, because measures of party identification are typically assessed using fully labeled 7-point scales, and measures of political efficacy are normally measured using 2- and 3-point scales. Controlling for question characteristics should reduce the strength of association between question content and reporting reliability, but the relation between topic and reliability is expected to maintain itself, even after controlling for characteristics of the survey questions used. Unfortunately, because of the nature of the confounding of question topic and question response format in the array of measures employed here, we were unable to further examine this hypothesis.

Our findings that levels of schooling were inversely related to the magnitudes of reliability confirmed our expectations. Such results have a relatively plausible interpretation linked to the convergence of schooling experiences and the requirements of survey attitude reporting. We find no systematically monotonic decline in attitude reporting reliability with age, although the oldest age group in some cases shows a significantly lower level of reliability. There is no support in these data for the hypothesis that random reporting of attitudes increases with age. On the contrary, our best estimates of the role of age in reporting reliability suggest that declines in reliability accompanying aging are nonmonotonic and primarily occur in the oldest age groups.

Although it is desirable to assess the reliabilities of extant survey questions, as we have noted, such analyses leave certain variables confounded. Further experimental research may be required to ascertain the extent of contributions of various topic and question characteristics to estimated reliability. However, given that long-term panel data are necessary for optimal reliability estimation (such as the 3-wave panels with 2-year reinterview periods of the 1950s and 1970s NES designs; see Alwin 1989), such large-scale experimentation will not be a simple or inexpensive endeavor. Still, such between-subjects experimental designs incorporating nested within-subjects panels would help clarify several of the issues raised here. Until such designs

are feasible, we encourage the type of quasi-experimental analysis of existing data, such as that employed here. This will permit the development of a tentative understanding of the link between question characteristics and reliability of measurement.

NOTES

1. In this sense classical true score theory is less restrictive than is sometimes suggested (e.g., Bohrnstedt 1970; Zeller and Carmines 1980).

2. We recognize that some attitude researchers use composite scores rather than single items to measure attitudes. Most, however, analyze individual items. We do not address the reliability of composite variables, but our analysis is relevant to this issue because the reliability of composite scores is directly a function of the reliability of the component items (see Greene and Carmines 1979).

3. The index of reliability is defined as the square root of reliability (see Lord and Novick 1968). Unfortunately, there is some confusion about this in the survey methodology literature. Groves (1989, p. 42), for example, confuses "reliability" with the "index of reliability," defining the latter in the way the psychometric literature defines the former.

4. See Smith (1984) for a thorough documentation of the debate and discussion that Converse's work on nonattitudes has stimulated.

5. Sherif and his colleagues devoted relatively little attention to the factors that determine the size of an individual's latitude of acceptance with regard to a particular attitude object. However, Eagly and Telaak (1972) found that the size of a person's latitude of rejection varied with ego-involvement. Persons who were more ego-involved in an attitude object typically had smaller latitudes of acceptance. Therefore, one might expect that individuals who are more ego-involved in an attitude object would be likely to evidence less random error in their attitude reports.

6. The claim that response scales with more scale points are preferable can also be justified by the often-made argument that longer response scales communicate more information than do shorter scales (Cox 1980; Garner 1960; Green and Rao 1970). Previous research has shown that respondents differentiate more between objects when offered response scales with greater numbers of categories (Bendig 1954; Garner 1960).

7. Converse's (1964) original argument about nonattitudes was based on analyses of the 1956 to 1960 National Election Panel Study, in which explicit "don't know" filters were used, presumably in order to reduce the number of nonattitude reports that were given. Converse's analysis of these items produced what he claimed was evidence of many nonattitudes being measured. Therefore, either (a) levels of nonattitudes would have been even higher had the filters been omitted, or (b) asking the filter question did not effectively remove random responders, or (c) Converse's method for estimating the prevalence of nonattitudes was inaccurate and exaggerated.

8. In 1978 the primary sampling unit specifications were changed from standard metropolitan statistical areas and counties to fit congressional district lines, but this change should have no appreciable effect on the representativeness of the full sample.

9. Because of limitations of space we have not here reproduced the exact questions, their response categories, our coding schema, or the estimates of reporting reliability. These are presented in Alwin and Krosnick (1989) and may be obtained on request from the authors.

10. Because of extreme skew in their marginal distributions, we excluded 6 of the GSS reinterview attitude measures.

11. Feeling thermometers are often thought of as having 100 or 101 scale points because the codes range from 0 to 100. In fact, in the NES studies the respondents are shown a card that labels nine specific scores (0, 15, 30, 40, 50, 60, 70, 85, and 100), along with explicit verbal labels for each of these 9 scale-points (see Weisberg and Miller, n.d.). Respondents giving scores in between these labelled scale-points, for example, a score of 20 or 55, are recorded, however, our analysis indicates that rarely more than 3% to 5% of respondents give a response other than the 9 labelled numeric options. Thus, for all intents and purposes, this is a 9-point scale. Our analysis of these data preserves the coding of finer gradations when they exist.

12. See Alwin (1989) for a detailed discussion of these models.

13. Because of the categoric nature of survey data, there is justifiable concern that analysis by LISREL may not be appropriate, either because it is commonly assumed that LISREL requires multivariate normality or because such continuous-variable models are felt to be inappropriate for categoric data. We, thus, estimated our models using the generalized least squares and unrestricted least squares least-squares options in LISREL, as well as the maximum likelihood option. Results across the three estimation approaches were identical for our model. Alwin (1991) further subjected these data to estimation using least-squares estimates obtained using EQS (Bentler 1989). Specifically the GLS arbitrary-distribution-theory estimates were obtained using EQS. These GLS (ADT) estimates were either identical or quite similar for all variables used in this analysis. In addition, Alwin (1991) also estimated these models using LISCOMP (Muthén 1987) estimation based on polychoric correlations. Although differing in order of magnitude (the LISCOMP estimates were generally higher), Alwin (1991) reached the same conclusions using LISCOMP estimates as were reached using maximum-likelihood estimates from LISREL or GLS ADT estimates from EQS. We conclude on the basis of Alwin's (1991) investigation that the strategy of estimating reliability probably does not greatly affect one's conclusions.

14. Alwin and Krosnick (1989) presented the three reliability estimates, one for each timepoint, from the Wiley and Wiley (1970) model, but as indicated we analyze reliability estimate from only the second time.

15. MCA is formally equivalent to multiple regression with categoric predictors or multi-way ANOVA with no interactions. We wish to thank Willard Rodgers for access to his program for modifying the MCA program to obtain information for the statistical evaluation of MCA results. Specifically, the Rodgers MCA program provides p values for the statistical significance of the deviation of category means from the grand mean, which the Andrews et al. (1973) program does not.

16. Statistical tests on category differences in reliability are not technically appropriate, because questions have been neither randomly selected from some know universe of questions, nor are they independent in a sampling sense. We present information from the statistical test nonetheless in order to illustrate something about the relative magnitude of a particular relationship, not as a basis for generalizing to some known universe of questions.

17. For purposes of most analyses presented below, we combine the latter three categories: party, candidates, and ideology.

18. The GSS included one "party identification" question, but it was eliminated from the present analysis because of unresolvable coding errors.

19. In the analysis of topic differences in the 1980s NES, we excluded groups with fewer than two measures. Thus this test involved the comparison of the policy, party, and candidate measures.

20. The 14 agree-disagree questions are all in the 1950s and 1970s panels and virtually all deal with policy attitudes. Thus time between waves and topic are controlled by selection, and only the unadjusted means are presented in Table 3. On the other hand, the rating scales come from the 1980s as well as the earlier panels, and we must control for time between waves. There is also variation in the topics of these questions, and topic is thus controlled in this analysis.

21. It should be pointed out that the comparison of the 2- versus 5-category agree-disagree scales actually confounds topic and the offering of an explicit "don't know" option. All of the 2-category questions deal with efficacy and do not provide a "don't know" option, whereas the 5-category questions almost all deal with policy attitudes and offer an explicit "don't know" choice. Thus it is somewhat doubtful whether these contrasts provide an adequate test of the hypothesis at hand.

22. The adjusted means presented here control for time between reinterviews, but this has no effect on the results.

23. The "don't know" responses are deleted from the analyses reported in this article. In fact, in order to be included in the analysis for any of the results reported in this article a respondent had to have "nonmissing" data for all three timepoints. Thus if the explicit offering of a "don't know" option removes persons who would report randomly, then we would expect the remaining respondents to be more reliable in their reports of attitudes.

24. The GSS reinterview samples were too small to permit separate analysis of reliability by age.

25. As noted at the bottom of Table 6, in a few instances we excluded attitude questions from the analysis because in obtaining reliability estimates from LISREL, skewed distributions of responses tended to produce estimates that were unlikely or a model that would not converge. In order to remove this effect from the results presented here, we excluded such measures from a particular age or education category when it occurred.

REFERENCES

Abelson, R. P. 1972. "Are Attitudes Necessary?" Pp. 19-32 in *Attitudes, Conflict and Social Change*, edited by B. T. King and E. McGinnies. New York: Academic Press.

Achen, C. H. 1975. "Mass Political Attitudes and the Survey Response." *American Political Science Review* 69:1218-31.

Alwin, D. F. 1973. "Making Inferences from Attitude-Behavior Correlations." *Sociometry* 36:253-78.

———. 1988. "Structural Equation Models in Research on Human Development and Aging." Pp. 71-170 in *Methodological Issues in Aging Research*, edited by K. W. Schaie, R. T. Campbell, W. Meredith, and S. C. Rawlings. New York: Springer.

———. 1989. "Problems in the Estimation and Interpretation of the Reliability of Survey Data." *Quality and Quantity* 23:277-331.

———. 1991, August. "Information Transmission in the Survey Interview: Number of Response Categories and the Reliability of Attitude Measurement." Paper presented at the 1991 annual meetings of the American Sociological Association. Cincinnati, OH.

Alwin, D. F., and D. J. Jackson. 1979. "Measurement Models for Response Errors in Surveys: Issues and Applications." Pp. 68-119 in *Sociological Methodology 1980*, edited by K. F. Schuessler. San Francisco: Jossey-Bass.

Alwin, D. F., and J. A. Krosnick. 1989, May. "The Reliability of Attitudinal Survey Measures." Paper presented at the annual meetings of the American Association of Public Opinion Research. St. Petersburg, FL.

———. Forthcoming. "Aging, Cohorts, and the Stability of Socio-Political Orientations over the Life-Span." *American Journal of Sociology*.

Andrews, F. M. 1984. "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach." *Public Opinion Quarterly* 48:409-42.

Andrews, F. M. and A. R. Herzog. 1986. "The Quality of Survey Data as Related to Age of Respondent." *Journal of the American Statistical Association* 81:403-10.

Andrews, F. M., J. N. Morgan, J. A. Sonquist, and L. Klem. 1973. *Multiple Classification Analysis*. Ann Arbor, MI: Institute for Social Research.

Andrews, F. M. and S. B. Withey. 1976. *Social Indicators of Well-Being: Americans' Perceptions of Life Quality*. New York: Plenum Press.

Bem, D. J. 1972. "Self-Perception Theory." Pp. 1-62 in *Advances in Experimental Social Psychology*, Vol. 6, edited by L. Berkowitz. New York: Academic Press.

Bendig, A. W. 1953. "The Reliability of Self-Ratings as a Function of the Amount of Verbal Anchoring and the Number of Categories of the Scale." *The Journal of Applied Psychology* 37:38-41.

———. 1954. "Transmitted Information and the Length of Rating Scales." *Journal of Experimental Psychology* 47:303-8.

Bentler, P. M. 1989. *EQS—Structural Equations Program Manual*. BMDP Statistical Software, Inc. 1440 Sepulveda Boulevard, Suite 316, Los Angeles, CA.

Bishop, G. F., Oldendick, R. W., and Tuchfarber, A. J. 1983. "Effects of Filter Questions in Public Opinion Surveys." *Public Opinion Quarterly* 47:528-46.

Bohrnstedt, G. W. 1970. "Reliability and Validity Assessment in Attitude Research." Pp. 80-99 in *Attitude Measurement*, edited by G. F. Summers. Chicago: Rand McNally.

———. 1983. "Measurement." Pp. 70-121 in *Handbook of Survey Research*, edited by P. H. Rossi, J. D. Wright, and A. B. Anderson. New York: Academic Press.

Bohrnstedt, G. W. and T. M. Carter. 1971. "Robustness in Regression Analysis." Pp. 118-46 in *Sociological Methodology 1971*, edited by H. L. Costner. San Francisco: Jossey-Bass.

Bohrnstedt, G. W., P. P. Mohler, and W. Müller. 1987. "Editor's Introduction." *Sociological Methods & Research* 15:171-76.

Champney, H. and H. Marshall. 1939. "Optimal Refinement of the Rating Scale." *Journal of Applied Psychology* 23:323-31.

Cleary, T. A., R. L. Linn, and G. W. Walster. 1970. "Effect of Reliability and Validity on Power of Statistical Tests." Pp. 130-38 in *Sociological Methodology 1970*, edited by E. F. Borgatta and G. W. Bohrnstedt. San Francisco: Jossey-Bass.

Converse, P. E. 1964. "The Nature of Belief Systems in the Mass Public." Pp. 206-61 in *Ideology and Discontent*, edited by D. E. Apter. New York: Free Press.

———. 1970. "Attitudes and Non-Attitudes: Continuation of a Dialogue." Pp. 168-89 in *The Quantitative Analysis of Social Problems*, edited by E. R. Tufte. Reading, MA: Addison-Wesley.

———. 1974. "Comment: The Status of Non-Attitudes." *American Political Science Review* 68:650-60.

———. 1980. "Rejoinder to Judd and Mulburn." *American Sociological Review* 45: 644-46.

Converse, P. E. and G. B. Markus. 1979. "Plus Ça Change . . . : The New CPS Election Study Panel." *American Political Science Review* 73:32-49.

Coombs, C. H. and L. C. Coombs. 1976. " 'Don't Know': Item Ambiguity or Respondent Uncertainty?" *Public Opinion Quarterly* 40:497-514.

Cox, E. P., III. 1980. "The Optimal Number of Response Alternatives for a Scale: A Review." *Journal of Marketing Research* 27:407-22.

Eagly, A. H. and K. Telaak. 1972. "Width of the Latitude of Acceptance as a Determinant of Attitude Change." *Journal of Personality and Social Psychology* 23:388-97.

Fazio, R. H., P. M. Herr, and T. J. Olney. 1984. "Attitude Accessibility Following a Self-Perception Process." *Journal of Personality and Social Psychology* 47:277-86.

Ferguson, L. W. 1941. "A Study of the Likert Technique of Attitude Scale Construction." *Journal of Social Psychology* 13:51-57.

Finn, R. H. 1972. "Effects of Some Variations in Rating Scale Characteristics on the Means and Reliabilities of Ratings." *Educational and Psychological Measurement* 32:255-65.

Garner, W. R. 1960. "Rating Scales, Discriminability, and Information Transmission." *The Psychological Review* 67:343-52.

Green, P. E. and V. R. Rao. 1970. "Rating Scales and Information Recovery – How Many Scales and Response Categories to Use?" *Journal of Marketing* 34:33-39.

Greene, V. L. and E. G. Carmines. 1979. "Assessing the Reliability of Linear Composites." Pp. 160-75 in *Sociological Methodology 1980*, edited by K. F. Schuessler. San Francisco: Jossey-Bass.

Groves, R. M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Heise, D. R. 1969. "Separating Reliability and Stability in Test-Retest Correlations." *American Sociological Review* 34:93-101.

Jahoda, M., M. Deutsch, and S. W. Cook. 1951. *Research Methods in Social Relations*. New York: Dryden Press.

Jenkins, G. D., Jr. and T. D. Taber. 1977. "A Monte Carlo Study of Factors Affecting Three Indices of Composite Scale Reliability." *Journal of Applied Psychology* 62:392-98.

Jöreskog, K. G. 1970. "Estimation and testing of simplex models." *British Journal of Mathematical and Statistical Psychology* 23:121-45.

———. 1974. "Analyzing Psychological Data by Structural Analysis of Covariances Matrices." Pp. 1-56 in *Measurement, Psychophysics, and Neural Information Processing*, edited by D. H. Kranz, R. C. Atkinson, R. D. Luce, and P. Suppes. San Francisco: Freeman.

Jöreskog, K. G., and D. Sörbom. 1986. *LISREL VI. Analysis of Linear Structural Relationships by Maximum Likelihood, Instrumental Variables, and Least Squares Methods*. Scientific Software, Inc. P.O. Box 536, Mooresville, Indiana 46158.

Judd, C. M., J. A. Krosnick, and M. A. Milburn. 1981. "Political Involvement and Attitude Structure in the General Public." *American Sociological Review* 46:660-69.

Judd, C. M. and M. A. Milburn. 1980. "The Structure of Attitude Systems in the General Public: Comparisons of a Structural Equation Model." *American Sociological Review* 45:627-43.

Kalton, G. and H. Schuman. 1982. "The Effect of the Question on Survey Responses: A Review." *Journal of the Royal Statistical Association* 145:42-73.

Kogan, N. 1961. "Attitudes Toward Old People in an Older Sample." *Journal of Abnormal and Social Psychology* 62:616-22.

Komorita, S. S. 1963. "Attitude Content, Intensity, and the Neutral Point on a Likert Scale." *Journal of Social Psychology* 61:327-34.

Komorita, S. S. and W. K. Graham. 1965. "Number of Scale Points and the Reliability of Scales." *Educational and Psychological Measurement* 25:987-95.

Krippendorff, K. 1970. "Bivariate Agreement Coefficients for Reliability of Data." Pp. 139-50 in *Sociological Methodology 1970*, edited by E. F. Borgatta and G. W. Bohrnstedt. San Francisco: Jossey-Bass.

Krosnick, J. A. and D. F. Alwin. 1989. "Aging and the Susceptibility to Attitude Change." *Journal of Personality and Social Psychology* 57:416-425.

Lehmann, D. R. and J. Hulbert. 1972. "Are Three-Point Scales Always Good Enough?" *Journal of Marketing Research* 9:444-46.

Lissitz, R. W. and S. B. Green. 1975. "Effect of the Number of Scale Points on Reliability: A Monte Carlo Approach." *Journal of Applied Psychology* 60:10-13.

Lord, F. M. and M. R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Madden, J. M. and R. D. Bourdon. 1964. "Effects of Variations in Scale Format on Judgement." *Journal of Applied Psychology* 48:147-51.

Markus, G. B. 1982. "Political Attitudes During an Election Year: A Report on the 1980 NES Panel Study." *American Political Science Review* 76:538-60.

Matell, M. S. and J. Jacoby. 1971. "Is There an Optimal Number of Alternatives for Likert Scale Items? Study I: Reliability and Validity." *Educational and Psychological Measurement* 31:657-74.

———. 1972. "Is There an Optimal Number of Alternatives for Likert Scale Items? Effects of Testing Time and Scale Properties." *Journal of Applied Psychology* 56:506-9.

Miller, D. P. and A. D. Swain. 1987. "Human Error and Human Reliability." Pp. 219-50 in *The Handbook of Human Factors* edited by G. Salvendy. New York: Wiley.

Miller, G. A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63:81-97.

Miller, W. E., A. H. Miller, and E. J. Schneider. 1980. *American National Election Studies Data Sourcebook, 1952-1978*. Cambridge, MA: Harvard University Press.

Moser, C. A. and G. Kalton. 1972. *Survey Methods in Social Investigation*. New York: Basic Books.

Murphy, G. and R. Likert. 1938. *Public Opinion and the Individual: A Psychological Study of Student Attitudes on Public Questions, with a Retest Five Years Later*. New York: Russell and Russell.

Muthén, Bengt. 1987. *LISCOMP: Analysis of Linear Structural Equations Using a Comprehensive Measurement Model*. Scientific Software, Inc. P.O. Box 536, Mooresville, IN 46158.

National Opinion Research Center. 1987. *General Social Surveys, 1972-87: Cumulative Codebook*. Chicago: Author.

Peters, D. L., and E. J. McCormick. 1966. "Comparative Reliability of Numerically Anchored versus Job-Task Anchored Rating Scales." *Journal of Applied Psychology* 50:92-96.

Rodgers, W. L., F. M. Andrews, and A. R. Herzog. 1989. "Quality of Survey Measures: A Structural Modeling Approach." Unpublished paper, Institute for Social Research, Ann Arbor, MI.

Rodgers, W. L. and A. R. Herzog. 1987a. "Interviewing Older Adults: The Accuracy of Factual Information." *Journal of Gerontology* 42:387-94.

———. 1987b. "Measurement Error in Interviews with Elderly Respondents." Unpublished paper, Institute for Social Research, Ann Arbor, MI.

Rugg, D. and H. Cantril. 1944. "The Wording of Questions." Pp. 23-50 in *Gauging Public Opinion*, edited by H. Cantril. Princeton: Princeton University Press.

Saris, W. E. 1988. *Variation in Response Functions: A Source of Measurement Error in Attitude Research*. Amsterdam: Sociometric Research Foundation.

Schuman, H. and S. Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments in Question Form, Wording and Context*. New York: Academic Press.

Sears, D. O. 1981. "Life-Stage Effects on Attitude Change, Especially Among the Elderly." Pp. 183-204 in *Review of Personality and Social Psychology*, Vol. 4, edited by S. B. Kiesler, J. N. Morgan, and V. K. Oppenheimer. Beverly Hills, CA: Sage.

Sherif, M., and C. I. Hovland. 1961. *Social Judgement: Assimilation and Contrast Effects in Communication and Attitude Change*. New Haven: Yale University Press.

Sherif, C. W., M. Sherif, and R. E. Nebergall. 1965. *Attitude and Attitude Change*. Philadelphia: W. B. Saunders.

Simon, H. 1977. *Models of Discovery*. Amsterdam: D. Reidel, Dordrecht.

Smith, T. W. 1980. "Inconsistent People." Report prepared for the Panel on the Measurement of Subjective Phenomena, National Academy of Sciences. GSS Technical Report #49. Chicago: National Opinion Research Center.

———. 1984. "Nonattitudes: A Review and Evaluation." Pp. 215-55 in *Surveying Subjective Phenomena*, Vol. 2, edited by C. F. Turner and E. Martin. New York: Russell Sage Foundation.

Smith, T. W. and B. L. Peterson. 1985, August. "The Impact of Number of Response Categories on Inter-Item Associations: Experimental and Simulated Results." Paper presented at the annual meetings of the American Sociological Association, Washington, DC.

Smith, T. W. and C. B. Stephenson. 1979. "An Analysis of Test/Retest Experiments on the 1972, 1973, 1974, and 1978 General Social Surveys." GSS Technical Report, No. 14, December, Chicago: National Opinion Research Center.

Symonds, P. M. 1924. "On the Loss of Reliability in Ratings Due to Coarseness of the Scale." *Journal of Experimental Psychology* 7:456-61.

Weisberg, H. F. and A. H. Miller. n.d. "Evaluation of the Feeling Thermometer: A Report to the National Election Study Board Based on Data from the 1979 Pilot Survey." Unpublished paper. Center for Political Studies, Institute for Social Research, Ann Arbor, MI.

Werts, C. E., K. G. Jöreskog, and R. L. Linn. 1971. "Comment on 'The Estimation of Measurement Error in Panel Data.' " *American Sociological Review* 36:110-13.

Wicker, A. W. 1969. "Attitudes versus Actions: The Relationship of Verbal and Overt Behavioral Responses to Attitude Objects." *Journal of Social Issues* 25:41-78.

Wiley, D. E. and J. A. Wiley. 1970. "Estimating Measurement Error Using Multiple Indicators and Several Points in Time." *American Sociological Review* 35:112-17.

Zaller, J. 1988, September. "Vague Minds vs. Vague Questions: An Experimental Attempt to Reduce Measurement Error." Paper presented at the annual meetings of the American Political Science Association, Washington, DC.

Zeller, R. and E. E. Carmines. 1980. *Measurement in the Social Sciences*. New York: Cambridge University Press.

Duane F. Alwin is a professor of sociology and a program director at the Survey Research Center of the Institute for Social Research, University of Michigan, where he works on problems of survey measurement (see the introductory article in this issue). In addition to his contributions to survey research methodology, Alwin has published in the American Sociological Review, American Journal of Sociology, Journal of Marriage and the Family, Sociology of Education,

and Journal of Personality and Social Psychology *on the topics of social change, the family, religion, education, human development and aging, and social psychological aspects of social inequality and stratification. He is a coauthor (with Ronald Cohen and Theodore Newcomb) of* Political Attitudes Over the Life-Span: The Bennington Women After Fifty Years *(Forthcoming, University of Wisconsin Press).*

Jon A. Krosnick is an associate professor of psychology and political science at Ohio State University. His M.A. and Ph.D. degrees from the University of Michigan are in psychology. Krosnick is the author of two books, as well as articles appearing in the Journal of Personality and Social Psychology, American Sociological Review, *and* American Political Science Review. *Krosnick's current research explores the dynamics of political attitudes and the cognitive processes underlying responses to survey questions.*