Effects," in N. Schwarz, and S. Sudman (eds.), *Context Effects in Social and Psychological Research*, New York: Springer Verlag, pp. 35–47.

Tourangeau, R., and Rasinski, K. (1988), "Cognitive Processes Underlying Context Effects in Attitude Measurement," *Psychological Bulletin*, 103, pp. 299–314.

Tourangeau, R., Rasinski, K., Bradburn, N., and D'Andrade, R. (1989a), "Carryover Effects in Attitude Surveys," *Public Opinion Quarterly*, 53, pp. 495–524.

Tourangeau, R., Rasinski, K., Bradburn, N., and D'Andrade, R. (1989b), "Belief Accessibility and Context Effects in Attitude Measurement," *Journal of Experimental Social Psychology*, 25, pp. 401–421.

Wyer, R., and Srull, T. (1989), *Memory and Cognition in Its Social Context*, Hillsdale, NJ: Erlbaum.

In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), Survey Measurement and Process Quality. New York: John Wiley, 1997.

CHAPTER 6

Designing Rating Scales for Effective Measurement in Surveys

Jon A. Krosnick
The Ohio State University

Leandre R. Fabrigar
Queen's University at Kingston

6.1 INTRODUCTION

Rating scales are omnipresent in contemporary surveys measuring subjective phenomena such as attitudes and beliefs.[1] Beginning with the pioneering work of Thurstone and Chave (1929), Likert (1932), and their contemporaries, a long line of research has documented the utility of rating scales for this purpose. As such, rating scales play central roles in empirical studies by psychologists, sociologists, political scientists, economists, and many other social scientists. Yet remarkably, research methods and questionnaire design textbooks rarely provide specific, practical advice about how best to construct rating scales for particular purposes.

Such advice seems potentially useful because designing a rating scale entails making many different decisions. One must establish how long a scale will be and whether it will include a midpoint. One must decide whether to include verbal labels on all points or to label some with numbers only. If one chooses to use verbal labels, decisions must be made about precisely which words to

[1] Although the term "scale" is often used to refer to a battery of items measuring a single construct, we use the term "rating scale" in this chapter to refer to the response options used in a single question employing a rating format.

Survey Measurement and Process Quality, Edited by Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, Trewin.
ISBN 0-471-16559-X © 1997 John Wiley & Sons, Inc.

use. If one uses numeric labels, they must be chosen as well, and a no-opinion response option can be either included or omitted. For lack of an alternative approach, most questionnaire designers make these decisions by relying on their intuition, so different investigators sometimes end up employing very different procedures.

Most survey methodologists are unaware, however, that there is a very large empirical literature, scattered across eight decades and various different social sciences, that offers scientific bases for making these design decisions. At present, there is much work ongoing that will synthesize this literature, as well as research on many related topics of interest to questionnaire designers (e.g., see Krosnick and Fabrigar, forthcoming). In this chapter, we preview some of what is being uncovered in that literature. Specifically, we focus on three principal issues of interest in rating scale design: number of scale points, verbal versus numeric labeling, and inclusion of no-opinion options. We begin with a brief discussion of the criteria we use to assess the quality of a rating scale's design: reliability and validity. We then review some of the literature on each of our three principal topics.

6.2 EVALUATING DATA QUALITY

Research methods textbooks in the social sciences routinely point to two fundamental aspects of the quality of a measure: reliability and validity. Because there are many excellent and extensive published discussions of these issues elsewhere, there is no need to define or refine these criteria here. However, it does seem worthwhile to note the inherent strengths and weaknesses in methods used to assess item reliability and validity in surveys in order to help us make sense of the evidence we consider below that may at times appear to be contradictory.

6.2.2 Reliability

In conducting surveys to measure attitudes, we presume that attitudes are latent constructs that reside in the minds of individuals. No single question can tap such a construct perfectly, because answers to any question will be a function of a variety of extraneous factors other than the attitude itself. But the construct nonetheless resides in the individual's mind, and the question provides a reading of it, subject to the influence of the other forces. A question's reliability can be gauged by assessing the consistency of its readings, either by the consistency of results obtained when the same person is asked the same question on multiple occasions (what we will call *longitudinal reliability*), or by the consistency of results obtained from asking the person a series of different questions intended to tap the same attitude on one occasion (called *cross-sectional reliability*).

Each method has drawbacks. A lack of correspondence in answers to an item over time can be observed for either of two reasons: the measure may be unreliable, or the attitude itself may have changed during the intervening time interval. The shorter the time interval, the less likely attitude change is to occur,

but the more likely it is that the person will remember his or her answer to the question during its first administration and simply repeat it, thus artifactually inflating apparent reliability.

A series of questions intended to measure the same attitude on a single occasion may be perceived by a respondent as just that, and he or she may strive to provide answers that appear to be consistent with one another across the items. Or respondents may doubt that a researcher would intentionally ask a series of questions about exactly the same issue, so the respondents may attempt to infer fine distinctions between the questions and thus exaggerate differences between them. Given these drawbacks of each method, we should not expect every study to yield perfectly pure results. Rather, we should look for trends in findings obtained by a variety of different methods.

6.2.2 Validity

The validity of a measure refers to the accuracy with which it taps the construct of interest. Reliability and validity are related to one another causally, in that a measure with no reliability at all cannot be valid. However, a measure that is highly reliable may nonetheless be quite inaccurate. Its consistency may come instead from a fundamental insensitivity or from tapping a construct other than the one of interest.

The validity of survey measures can be assessed in a number of ways. For instance, what we will call *correlational validity* is the degree to which a given measure can predict other variables to which it should be related. Presumably, a stronger relation would indicate greater validity. So, for example, one attitude measure would be considered to be more valid than another if the former predicts attitude-relevant behavior better. Discriminant validity is the degree to which a measurement approach can differentiate between constructs that are presumed to be distinct from one another. Therefore, the validity of a rating scale is presumably greater if it is better able to detect differences in perceptions of different objects. Of course, the utility of these approaches to validity assessments depends upon the researcher's ability to correctly identify criteria that should be predicted by a measure and to identify objects that should be evaluated differently from one another. Consequently, it is again most useful to look for trends in results across large sets of studies that are heterogeneous in terms of design.

6.3 NUMBER OF SCALE POINTS

One important decision that must be made when constructing a rating scale is how many scale points to include, and a number of theoretical issues enter into this determination. It is important to distinguish between bipolar scales (i.e., scales reflecting two opposing alternatives with a clear conceptual midpoint) and unipolar scales (i.e., scales reflecting varying levels of some construct with

no conceptual midpoint and with a zero point at one end). Attitudes can be thought of as bipolar constructs, because they range from extremely positive to extremely negative, with neutral as a specific midpoint, representing neither positive nor negative. The amount of importance a person attaches to a particular attitude he or she holds is an example of a unipolar construct: it ranges from zero importance to some maximum level, and there is no precise midpoint.

There are various reasons to believe that more scale points will generally be more effective than fewer. This is because people's perceptions of their attitudes presumably range along a continuum of extremely positive to extremely negative. In order to translate a point on that continuum on to a categorical response scale, the set of points must presumably represent the entire continuum. A scale with only three options (e.g., favor, oppose, and neither) does not allow people to say that they favor something slightly. Thus, it is ambiguous as to whether a person with a slight leaning should select "favor" (which might imply stronger positivity than is the case) or "neither." A more refined scale with more points will presumably permit such moderate individuals to express their stands precisely and comfortably. Furthermore, the more scale points there are, the more a person can distinguish his or her attitude toward one object from his or her attitude toward another object. Similarly, more scale points permit a researcher to make more subtle distinctions among individuals' attitudes toward the same object. Thus, longer scales have the potential to convey more useful information.

On the other hand, using too many scale points may reduce the clarity of meaning of the response options. For scales with only a few options (e.g., the 3-point scale with favor, oppose, and neither), the meaning of these options is quite clear. But when the number of options becomes quite large (e.g., a 101-point thermometer scale), the meaning of any particular point is less precise. This may lead to less consistency within and between individuals concerning the meanings they attach to particular response options. In addition, including too many response options may make it more difficult for respondents to decide where they fall on a scale and thereby encourage respondents to shortcut the effort they expend via satisficing (Krosnick, 1991).

Thus, the optimal scale length would seem to be a moderate one. But just exactly how long should it be? With bipolar constructs, one might imagine that the scale should be set up to represent cognitive categories likely to capture the differentiations people make naturally. So with regard to attitudes, for example, people might be inclined to think of their liking of an object as being either slight, moderate, or substantial. If we think of three such points to the left of a midpoint on an attitude scale and three such points to the right, a 7-point scale appears optimal. It is also possible that people might only naturally distinguish between slight and substantial leaning to one side or the other, thus implying that a 5-point scale may be optimal.

In the case of unipolar scales, using too few scale points probably compromises information gathered, and too long of a scale probably compromises

clarity of meaning. But it is hard to anticipate what the optimal length will be. It seems likely that people can readily conceive of zero, a slight amount, a moderate amount, and a great deal along any unipolar continuum. And perhaps they are comfortable making slightly finer distinctions, so the optimal scale length might be expected to fall between 4–7 points.

6.3.1 Reliability

Much of the empirical research exploring the effect of scale point number on measurement quality has investigated its influence on reliability. This research has used a variety of approaches, including secondary analyses of existing data and direct experimental comparisons.

Although some studies have failed to find a relation between the number of points on bipolar scales and cross-sectional reliability (Bendig 1954a), most studies indicate that such reliability is greatest for scales with approximately 7 points. For example, Masters (1974) experimentally compared 2-, 3-, 4-, 5-, 6-, and 7-point scales and found that reliability increased up to 4 points and leveled off thereafter. Birkett (1986) experimentally compared 2-, 6-, and 14-point scales and found that 6-point scales had the highest reliability. Similarly, Komorita and Graham (1965) found 6-point scales to be more reliable than 2-point scales. In one study that examined longitudinal reliability, 7- to 9-point scales appeared to be more reliable than shorter ones (Alwin and Krosnick, 1991).

Investigations of cross-sectional reliability in unipolar scales suggest that the optimum number of scale points is between 5 and 7 points. Although some studies failed to find systematic relations between scale point number and reliability (Bendig, 1953; Matell and Jacoby, 1971; Peterson, 1985), other studies found that cross-sectional reliability is greater for 4-point than 2-point scales (Watson, 1988), 5-point than 7- or 11-point scales (McKelvie, 1978), and 5- to 7-point than 3- or 9-point scales. In an investigation of longitudinal reliability (Jacoby and Matell, 1971; Matell and Jacoby, 1971), 7- and 8-point scales had greater reliability than scales with 3 to 6 points or scales with 9 to 19 points.

6.3.2 Validity

A substantial amount of research has also examined the effect of scale point number on the validity of measurement. Much of this research has used computer simulations to examine how transforming data from continuous representations of relations to representations with discrete scale points distorts known patterns of data. With a few exceptions (Martin, 1973, 1978), these simulations suggest that distortion in data decreases as the number of scale points increases, but that this improvement is relatively modest beyond 5 to 7 points (Green and Rao, 1970; Lehmann and Hulbert, 1972; Ramsay, 1973).

Studies of correlational validity are generally consistent with the above-mentioned reliability patterns, even though some studies have produced contradictory evidence (e.g., Smith and Peterson, 1985). For instance, Matell

and Jacoby (1971) found greater correlational validity for 7- and 8-point unipolar scales than for shorter or longer scales. Similarly, Rosenstone *et al.* (1986) found that 5-point bipolar scales predicted conceptually related variables better than 3-point bipolar scales.

Another set of studies relevant to assessing the effect of scale point number on validity has examined the susceptibility of rating scales to context effects (Wedell and Parducci, 1988; Wedell *et al.*, 1990). As expected, the effect of context on ratings of different stimuli along various dimensions of judgment (e.g., happiness of faces, size of geometric objects, happiness of life events) varies as a function of scale point number. Although results depend on the nature of the judgment being made, context effects generally weaken as the number of scale points increases, but these reductions are relatively modest beyond 7 scale points.

Another criterion for evaluating scales is the amount of information they yield about differences between attitudes within and across respondents. Needless to say, scales that are too short cannot reveal much about the distinctions a person makes among a large set of objects. Consistent with this notion, a number of studies showed that longer scales conveyed more useful information up to 7–9 points (Bendig, 1954b; Bendig and Hughes, 1953; Garner, 1960), and information transfer appears to decrease for scales of 12 points or longer (McCrae, 1970a, 1970b). Interestingly, the proportion of scale points used stays approximately constant across scales of between 4 and 19 points, but the additional number of points used on scales longer than 7 points seems not to convey additional substantive information (Matell and Jacoby, 1972).

6.3.3 Magnitude Scaling

A rather sizable literature has evolved proposing a radically different approach to rating scale construction under the label of "magnitude scaling" (e.g., Lodge, 1981). This approach involves the use of scales of infinite length, thus not restricting respondents in the extent to which they can differentiate among objects. For example, people could be shown a line four inches long and could be told that this represents the amount they like oranges. Then, they could be asked to draw another line indicating how much they like apples, with shorter lines indicating less liking and longer lines indicating more liking. In addition to the visual mode, this sort of scaling can be done by having people adjust the pitches of tones they hear or squeeze dynamometers to report comparative judgments in terms of the amount of pressure they apply.

As appealing as this method is in principle, however, it is difficult to administer practically, especially in typical survey settings. Also, magnitude scaling only reveals the ratios among stimuli in terms of their placement on an evaluative dimension, not the absolute levels of people's placement of them. That is, you can learn that a person likes apples twice as much as oranges, but you cannot tell whether apples or oranges fall on the positive or the negative side of his or her attitude continuum. Furthermore, a number of studies that compared the reliability and validity of magnitude estimation data with that obtained from traditional, categorical rating scales showed the latter to be superior (e.g., Kaplan *et al.*, 1979; Miethe, 1985). Thus, at the moment, it appears that the traditional approach is preferable.

6.3.4 Midpoints

Another important design issue is whether to include a middle alternative or not. On bipolar dimensions, a middle alternative can take a number of forms, including "neither" on a favor/oppose or agree/disagree or like/dislike scale, or a "status quo" endorsement on a scale representing change, ranging from a large increase to a large decrease (e.g., in U.S. defense spending). On a unipolar scale, a midpoint presumably represents a moderate position, though its meaning is not as precise as is the case for the bipolar dimensions.

Use of a middle alternative on a bipolar dimension can be justified if one believes that some individuals truly have neutral positions and that forcing them to respond in one direction or the other will add to measurement error. However, many people might lean slightly in one direction or the other but might select an offered midpoint because it provides an easy choice that requires little effort and is easy to justify. One scenario by which this might occur was proposed by Krosnick (1991).

According to Krosnick (1991), the cognitive tasks required of survey respondents are typically quite burdensome if people attempt to be fully diligent (a behavior he called survey *optimizing*). Consequently, individuals may sometimes look for ways to avoid expending this effort while maintaining the appearance of answering responsibly (a response behavior he called survey *satisficing*). According to Krosnick (1991), a respondent who wishes to satisfice will look for a cue in the question suggesting how to do so. If no such cue is obvious, then a respondent may choose to optimize instead.

Among various other cues, Krosnick (1991) pointed to scale midpoints. Specifically, he argued that when such points represent the status quo (e.g., "keep government spending on defense at present levels"), respondents will find them easy to defend if pressed by an interviewer. Consequently, offering a midpoint may discourage people from taking sides and may encourage them to satisfice, whereas if no midpoint is offered, respondents might optimize instead. Thus, offering a midpoint may forego collection of useful data.

Empirical research in this area has documented a variety of effects of offering scale midpoints. First, people seldom spontaneously offer midpoint responses when they are not legitimated, but substantial proportions select them when included explicitly in questions (Ayidiya and McClendon, 1990; Bishop, 1987; Kalton *et al.*, 1980; Schuman and Presser, 1981). Attraction to midpoints when offered is sometimes greatest among respondents who are lower in cognitive skills (Narayan and Krosnick, 1996) and lower in the personal importance of the topic involved (Krosnick and Schuman, 1988; Schuman and Presser, 1981, pp. 173–175; though see Stember and Hyman, 1949–1950). But although one might

imagine that such people flip mental coins to select responses when no middle alternative is offered, the distributions of opinions offered by these individuals often depart significantly from a flat shape (Ayidiya and McClendon, 1990; Bishop, 1987; Kalton *et al.*, 1980; Schuman and Presser, 1981). Thus, their responses do not appear to be haphazard.

Some additional studies have examined the effect of midpoint presence on reliability and validity, but with contradictory results. For example, Alwin and Krosnick's (1991) and Andrews's (1984) secondary analyses suggest that scales with midpoints were less reliable than those without. Experimental studies, however, have generally failed to find any consistent pattern of increases or decreases of reliability according to the presence or absence of midpoints (Masters, 1974; Jacoby and Matell, 1971; Matell and Jacoby, 1971). Although Kalton *et al.* (1980) found that including midpoints had no influence on correlational validity (as gauged by associations between attitudes and demographics), Schuman and Presser (1981) found that associations among attitudes were strengthened when middle alternatives were included, and Stember and Hyman (1949–1950) found that the effect of interviewer bias decreased with the inclusion of a midpoint. Thus, including a midpoint may at times increase data quality and at other times not.

6.3.5 Ease of Administration

Another criterion with which to assess the effectiveness of various scale lengths is the ease and success with which they can be administered to respondents. Matell and Jacoby (1972) found that it takes longer to administer scales of longer lengths, but sizable increases in time only occur after scales reach 13 points or longer. Of the studies that examined rates at which people properly completed all rating scales in a set, one found better completion rates for 4-point scales than 2-point scales (Ghiselli, 1939), but the other found no difference between 3- and 7-point scales (Smith and Peterson, 1985).

6.3.6 Conclusion

Taken as a whole, this research suggests that the optimal length of a rating scale is 5 to 7 points, because scales of this length appear to be more reliable and valid than shorter and longer scales. With respect to midpoints, it is less clear whether researchers should include midpoints. Evidence on validity is mixed, and the theory of satisficing suggests that including midpoints may decrease measurement quality. Therefore, we look forward to further studies on this matter to yield a clear resolution. In the meantime, however, it seems sensible to include midpoints when they are conceptually demanded (e.g., in a question about whether defense spending should be increased, decreased, or kept the same), and steps should be taken elsewhere in the questionnaire to minimize the likelihood of satisficing (see Krosnick, 1991).

6.4 LABELING SCALE POINTS

Another important decision in the construction of scales is whether to label all scale points with words or to label some scale points with numbers only. Obviously, in order for any rating scale to have meaning, it is necessary to at least label the endpoints of a scale. But researchers can decide not to label other scale points, and many routinely make this choice.

Using scales with only the endpoints labeled verbally presumably has two major advantages. First, numeric values may be more precise than verbal labels, which may suffer from the inherent ambiguity of language. Second, numeric scale values are presumably easier for respondents to hold in memory (e.g., during telephone interviews) than more complex verbal labels. Thus, responding to scales with only the endpoints verbally labeled may be less cognitively demanding than scales that are fully labeled.

However, there are also reasons to expect that verbally labeling all scale points might improve data quality. Because people rarely express complex conceptual meaning in everyday conversation via numbers, verbal labels might be a more natural (and therefore easier) method for respondents to express themselves. Additionally, numbered scale points have no inherent meaning, other than to suggest equal divisions between concepts established by verbal labels. Thus, including labels on all scale points could help to clarify the meaning of scale points and thereby increase the ease with which people can make reports and the precision of them.

6.4.1 Reliability

A number of empirical studies have compared fully labeled and partially labeled scales in terms of reliability. Studies of cross-sectional reliability found it to be unaffected by verbal labels (Finn, 1972; Madden and Bourdin, 1964) or to be lower for scales with more verbal labels (Andrews, 1984). However, studies of longitudinal reliability have consistently found it to be increased by verbal labels (Alwin and Krosnick, 1991; Krosnick and Berent, 1993; Zaller, 1988). Furthermore, these improvements in reliability were most pronounced among people with low to moderate education, just the individuals who can presumably benefit most from greater clarity (Krosnick and Berent, 1993). And using verbal labels to make the endpoints of a rating scale seem farther apart also increases reliability (Bendig, 1955). Thus, verbal labels seem to be an asset in this regard.

6.4.2 Validity

Studies of verbal labels also indicate that they increase the validity of obtained data. Krosnick and Berent (1993) and Dickinson and Zellinger (1980) found stronger correlations between attitudes and other variables when the former were measured using more verbal labels, especially among respondents with

relatively little formal education. In their secondary analyses, Andrews (1984) and Alwin and Krosnick (1991) found greater true score variance in fully labeled items than in partially labeled items, indicating greater validity of the former. Raters of the same objects tend to agree more when the rating scales employed have more verbal labels (Barrett *et al.*, 1958; Bendig, 1953; Peters and McCormick, 1966). Furthermore, more differentiation between different rating dimensions and different objects is apparent when more verbal labels are employed (Barrett *et al.*, 1958; Bendig and Hughes, 1953; Bernardin *et al.*, 1976). And finally, ratings are less susceptible to context effects when more verbal labels are employed (Wedell *et al.*, 1990). Thus, much evidence suggests that fully labeling scales improves measurement validity.

6.4.4 Respondent Satisfaction

A number of studies that have assessed respondent satisfaction found that most people preferred to use rating scales with more verbal labels (Dickinson and Zellinger, 1980; Wallsten *et al.*, 1993; Zaller, 1988) and believed such scales to be more valid measurement instruments (Dickinson and Zellinger, 1980).

6.4.4 Selecting Verbal Labels

Although fully labeled scales appear to be more reliable and valid than partially labeled scales, the benefits of verbal labeling are obviously contingent on selecting appropriate labels. If verbal labels are to be useful, they must have reasonably precise meanings for respondents. It is also important that the labels one chooses reflect relatively equal intervals along a continuum, particularly if an analyst is to capture all variance in the latent construct and plans to treat the results as an interval-level variable in statistical analysis.

A number of studies have investigated the meanings that people attach to verbal labels reflecting liking, amount, frequency, and likelihood—four dimensions routinely asked about in surveys. These studies have generally asked respondents to assign a numerical value (e.g., on a scale from 0 to 100 or via magnitude scaling) to various verbal quantifiers (e.g., excellent, good, fair). Many such labels seem to have meanings that are very stable across different samples and consistent over periods of many years and across different methods of rating (see Table 6.1 for an illustration of some results). Longitudinal reliability for these scale values has also been found to be quite high, with test-retest correlations frequently greater than .90. Thus, it is possible to find verbal labels with sufficiently precise meanings to construct useful scales.

Some researchers have looked at this literature and emphasized the imprecise nature of verbal labels suggested by research findings (e.g., Pepper, 1981). And it is certainly true that the meanings of some terms can change significantly depending upon various factors. For example, the absolute meaning of a frequency quantifier can change depending on the type of event being described

Table 6.1 Scale Values for Verbal Labels Assessing Liking

Verbal label	Myers and Warner (1968)	Wildt and Mazis (1978)	French-Lazovik and Gibson (1984)	Stone and Schkade (1991)
Excellent	93	91	91	99
Very good	78	82	80	
Good	67	70	58	73
Fair	43	49		48
Poor	21	17	16	23
Very poor	12	5	11	1

(e.g., earthquakes happening once a year might be described as "often," whereas going to the movies once a year might be described as "rarely"). Furthermore, people interpret the meanings of labels in the context of the other labels offered in a scale, presuming that they are intended to be equally spaced from one another. And different social groups sometimes interpret labels in systematically different ways (see, e.g., Schaeffer, 1991). However, we do not view this evidence with quite as much concern as do some other observers; although there is random and systematic variation in the meanings of verbal labels to respondents, many labels appear to have sufficiently universal meanings to be very useful for improving attitude measurement.

6.4.5 Selecting Numeric Labels

Even if a scale is fully verbally labeled, one may nonetheless be tempted to include numbers as well, partly to facilitate coding afterwards. Interestingly, on self-administered questionnaires, these numbers might have effects, just as numbers may affect ratings on partially labeled scales. This is because respondents sometimes use the specific numeric values to infer the meanings of scale points or verbal labels. For example, when respondents were asked how successful their lives have been, using an 11-point scale ranging from 0 to 10 with the endpoints labeled "not at all successful" and "extremely successful," approximately 34 percent selected a value ranging from 0 to 5 (see Schwarz and Hippler, 1991; Schwarz *et al.*, 1991). In contrast, when the same endpoints were used for an 11-point scale with numeric values ranging from -5 to $+5$, only 13 percent selected one of the logically equivalent values ranging from -5 to 0. This difference apparently occurred because respondents inferred that the label "not at all successful" meant the absence of success when paired with the number "0" but meant the presence of failure when paired with the number "-5."

Consequently, if numbers are to be used on rating scales, they should be selected carefully to reinforce the intended meaning of the scale points.

6.4.6 Conclusion

It seems that fully labeled scales are more reliable and valid than partially labeled scales. Based on this, we recommend that researchers verbally label all scale points when it is practical to do so and avoid the use of numbers alone. However, researchers should be certain that they select labels that have relatively precise meanings for respondents and that reflect equal intervals along the continuum of interest. This should be done by using labels that have been previously scaled and are known to possess good psychometric properties. When such labels are not available, researchers should conduct pretesting to identify appropriate labels. Because numeric values can alter the meaning of labels, researchers should probably avoid using them altogether and simply present verbal response options alone.

6.5 NO-OPINION FILTERS

When we ask attitude questions in surveys, we usually presume that respondents' answers reflect opinions that they previously had stored in memory. And if a person had not stored a pre-existing opinion about the precise attitude object of interest, the question itself presumably prompts him or her to draw on relevant stored information in order to concoct a reasonable, albeit new, evaluation of the object. Consequently, whether based upon a pre-existing attitude or a newly formulated one, responses presumably reflect the individual's orientation, favorable or unfavorable, toward the object.

6.5.1 The Non-Attitude Hypothesis

Converse (1964) proposed a very different possibility. He argued that respondents sometimes simply answer attitude questions randomly in surveys. His principal evidence was correlations between reports of the same attitudes made by a panel of respondents who were interviewed first in 1956, again in 1958, and finally in 1960. These correlations were remarkably weak: across eight policy issues, tau-betas ranged from approximately .28 (for federal housing) to approximately .47 (for school desegregation) and averaged .37. Because these correlations did not become notably weaker as the time interval between measurements increased from two years to four years, he concluded that no true attitude change had occurred. Furthermore, the distributions of opinions that the respondents expressed were remarkably consistent over the four-year period, which reinforced this same conclusion.

What, then, was responsible for the weak over-time correlations between these attitude reports if not attitude change? For some issues, Converse (1964) asserted, large portions of Americans had no pre-existing opinions but felt pressure to report attitudes in the surveys, even when they had absolutely no relevant information with which to formulate meaningful judgments. The mere presence of an interviewer and his or her persistence at asking questions presumably indicated to these respondents that they were expected to have opinions and that the researcher needed to know what those opinions were. Because respondents preferred not to look foolish by having to admit ignorance on numerous topics, Converse argued, respondents concocted answers even when they had no knowledge at all on which to base them.

6.5.2 Other Supportive Evidence

In fact, a great deal of other research, done both before and after Converse's (1964) chapter was published, is consistent with his conclusions. For example, one set of studies found comparably low levels of over-time consistency in attitude reports (e.g., Jennings and Markus, 1984; Jennings and Niemi, 1978). And latent class analyses of over-time data indicated that an average of between 50 and 70 percent of respondents had nonattitudes on any given public policy issue (Brody, 1986; Taylor, 1983).

Other research supporting the nonattitudes conclusion showed that attitude reports usually predict reports of other attitudes very weakly (for a review, see Kinder and Sears, 1985) and often predict relevant behavior rather poorly (Wicker, 1969). Furthermore, the results of studies showing strong attitude-behavior consistency may actually be artifactual consequences of correlated measurement error (Budd, 1987; Budd and Spencer, 1986). Still other supportive findings were reported by Gill (1947), Hartley (1946, pp. 26-31), Ehrlich and Rinehart (1965), Schuman and Presser (1981), Bishop *et al.* (1986), and others showing that people often report attitudes toward fictitious objects or objects so obscure that people are extremely unlikely to have heard of them.

Converse's (1964) interpretation of his initial evidence has received a great deal of criticism by observers who have claimed that other readings are equally plausible. Similarly, the evidence we just described as consistent with the nonattitudes hypothesis can also be interpreted in ways that are quite different (see Krosnick and Fabrigar, forthcoming). Thus, this literature does not provide definitive support for the nonattitude notion. However, its consistency with that hypothesis has led researchers to explore means by which nonattitude reports might be prevented.

6.5.3 Types of No-Opinion Filters

In fact, the only method that has been considered empirically to any significant degree is no-opinion filtering. The notion here is that respondents may report nonattitudes partly because the form of survey questions encourages them to do so. As Schuman and Presser (1981, p. 299) pointed out, respondents generally "play by the rules of the game," meaning that they choose among the response

alternatives offered by a closed-ended question rather than offering reasonable answers outside the offered set. If a question does not explicitly include a "don't know" (DK) or "no-opinion" (NO) option, that might imply to respondents that they are expected to have opinions and therefore encourage them to report nonattitudes. Thus, when such a response option is explicitly legitimated, significantly larger proportions of respondents would be expected to admit having no opinion.

Many studies have reported evidence consistent with this expectation (e.g., Ehrlich, 1964; Bishop *et al.*, 1980a, 1980b, 1982, 1983; 1986; Schuman and Presser, 1979, 1981; Schuman and Scott, 1989). Furthermore, filters can be phrased in various different ways, and these different approaches yield understandably different results. Quasi-filters involve simply including an explicit DK or NO response option at the end of a question asking respondents about their attitudes. So, for example, a question might ask: "Do you favor or oppose legalized abortion, or don't you have an opinion on this?" In contrast, full filters involve a question that precedes the attitude measure, inquiring about whether the respondent has an opinion on the issue. Only if he or she answers affirmatively is the attitude question then asked. Full filters attract significantly more DK responses (Bishop *et al.*, 1983; Schuman and Presser, 1981).

Full filters can be phrased in ways that vary in terms of their apparent encouragement and legitimation of DK responses, and this variation seems to be consequential as well (Bishop *et al.*, 1983). For example, one could ask "Do you have an opinion on this or not?" or "Have you been interested enough in this to favor one side over the other?" or "Have you thought much about this issue?" or "Have you already heard or read enough about it to have an opinion?" By focusing on a person's interest in, thoughts about, or exposure to information on a topic, the latter three filters presumably make it easier for respondents to admit that they have not considered the topic previously and therefore have no opinion on it. Consistent with this presumption, Bishop *et al.* (1983) found that filters can be ordered in terms of their strength: the blunt filter gathered slightly fewer DK responses than the "interested" filter, which gathered fewer DKs than the "thought" and "heard or read" filters.

6.5.4 The Meaning of No-Opinion Responses

This evidence is encouraging about the ability of no-opinion filters to prevent nonattitude reporting. However, this conclusion presumes that DK responses mean that respondents truly have no opinion and would guess purely randomly or would provide equally meaningless answers via some other method if forced to offer an opinion. But is this necessarily the case?

In fact, people may provide DK responses for many reasons other than having no attitudes at all (e.g., see Dunnette *et al.*, 1956; Edwards and Ostrom, 1971; Klopfer and Madden, 1980; Krosnick, 1991; Smith, 1984). First, DK responses may reflect ambivalent attitudes, which may be especially difficult to report when no scale midpoint is offered to respondents. Second, DK

responses could occur when a person has only neutral thoughts or feelings about an object but no neutral response option is offered (e.g., in an agree/disagree question). Third, selection of a DK alternative may reflect that the respondent does not understand the question being asked. Fourth, respondents may know approximately where they fall on an attitude scale (e.g., around 6 or 7 on a 1 7 scale), but because of ambiguity in the meaning of the scale points, they may be unsure of exactly which to choose. Finally, DK responses may reflect satisficing: even though respondents may have preformed opinions or information with which to concoct a reasonable opinion on the spot, lack of motivation or ability to do so may lead these individuals to choose the DK option in order to avoid doing the mental work involved (Krosnick, 1991).

Many studies suggest that don't know or no-opinion responses can have all of these different meanings (Coombs and Coombs, 1976; Dunnette *et al.*, 1956; Feick, 1989; Smith, 1984). Furthermore, some studies suggest that ambivalence is more often the cause of DK responses than nonattitudes (e.g., Ehrlich, 1964). Thus, if respondents who say "no-opinion" were forced to answer substantively instead, their responses might well not be primarily random noise.

6.5.5 Do No-Opinion Filters Improve Data Quality?

Given this backdrop, an interesting question here is whether offering a no-opinion option reduces the amount of random variation in the attitude reports obtained. Of course, nonattitude reports could be based upon systematic response biases, which would not necessarily increase random variation. But if Converse (1964) was correct that some nonattitude reports are purely random, then this randomness should weaken relations between variables. A variety of methods has been used to address this issue, and all of this work has produced either mixed evidence or clear disconfirmation of the random nonattitudes perspective.

For example, Gilljam and Granberg (1993) asked survey respondents three different questions tapping attitudes toward building nuclear power plants. The first of these questions included a no-opinion filter, and 15 percent of respondents selected it. The other two questions, asked later in the interview, did not include no-opinion filters, and only 3 and 4 percent of respondents, respectively, volunteered "don't know" responses to them. Thus, the majority of respondents who initially said "don't know" offered opinions on the later two questions. At issue, then, is whether these later responses reflected real opinions or were nonattitudes.

To address this issue, Gilljam and Granberg (1993) examined two indicators: the strength of the correlation between these two latter attitude reports, and their ability to predict people's votes on an actual nuclear power referendum in a subsequent election. The correlation between answers to the latter two items was .41 ($p < .001$) among individuals who said "don't know" to the first item, as compared to a correlation of .82 ($p < .001$) among individuals who answered the first item substantively. Similarly, answers to the second two items

correctly predicted an average of 76 percent of subsequent votes by people who initially said "don't know," as compared to a 94 percent accuracy rate among individuals who answered the first item substantively.

Thus, the nonattitudes perspective received mixed support, in that the filter apparently separated out people whose opinions were, on average, of less quality than others' opinions but were not purely random.

Two other nonexperimental studies taking a different investigative approach produced similarly mixed evidence. Andrews (1984) and Alwin and Krosnick (1991) meta-analyzed the correlates of the amount of random measurement error in numerous survey items, some of which included no-opinion filters and others of which did not. Andrews (1984) found that the amount of random error was significantly less when a no-opinion filter was included than when it was not, but Alwin and Krosnick (1991) found just the opposite.

One set of studies producing a clear challenge to the NO filter notion gauged changes in associations between target items and other variables. If offering a NO option leads some respondents to select it and thereby prevents them from providing poor quality data, then associations between variables should become stronger. Bishop *et al.* (1979) did a secondary analysis of existing surveys that had asked similar questions of different national samples in either filtered or unfiltered forms. These authors found a trend toward strengthened associations for filtered rather than unfiltered forms, but their comparisons involved confounding of time, survey administration form, question context, and more. In Schuman and Presser's (1981) 19 experiments on no-opinion filtering, no meaningful shifts in correlations occurred in the vast majority of cases. Two cases did turn up significant association shifts, but neither was consistent with Converse's expectations in terms of the directions of the shifts. Furthermore, Schuman and Presser (1981) found no cases in which relations between attitudes and respondents' education, interest in politics, age, or gender were altered by filtering. Thus, this evidence is clearly inconsistent with the notion that no-opinion filters improve the quality of data obtained.

A final pair of studies explored the effect of experimental variations in the presence or absence of NO filters on reliability. It would seem necessary for reliability to be increased by filters in order for the nonattitudes perspective on filters to be validated. However, using cross-sectional data, McClendon and Alwin (1993) found no greater reliability when NO filters were included in questions than when they were not. And Krosnick and Berent (1990) found no significant change in the over-time consistency of attitude reports depending upon whether no-opinion filters were present or absent in questions.

6.5.6 An Alternative Perspective

Taken together, these studies call into question the notion that NO filters successfully remove nonattitudes from the data obtained by a survey question. As surprising as these results are from Converse's (1964) perspective, they are

quite understandable in light of the notion of survey satisficing (Krosnick, 1991). Offering a no-opinion option may not only fail to improve the quality of data obtained but may also forego collection of useful data. Specifically, such a response option might constitute a cue encouraging respondents who are otherwise disposed to satisfice to do so by saying "don't know." If, instead, the no-opinion option was not offered and no other cue was apparent, these respondents might choose not to satisfice and would optimize instead. Thus, useful data could be collected from these individuals, and offering the no-opinion option would preclude obtaining these data.

If this view is correct, then attraction to a NO option when offered should be greatest under the conditions thought to foster satisficing: low respondent ability and motivation to optimize, and high task difficulty. Consistent with this reasoning, attraction is greatest among respondents with the lowest levels of cognitive skills (Bishop *et al.*, 1980a; Schuman and Presser, 1981; Schuman and Scott, 1989) and among people who consider an issue to be less personally important (Bishop *et al.*, 1980a; Schuman and Presser, 1981, pp. 142–143). However, these patterns can be viewed as consistent with Converse's (1964) nonattitudes perspective, because these individuals are precisely those who would seem most likely to be uninformed on an issue and to report nonattitudes. Therefore, additional evidence is needed before the meanings of these associations can become clear.

Although not optimal for testing the satisficing notion, some other studies have provided supportive evidence simply by correlating the frequency of no-opinion responses to a single question form with various aspects of task difficulty. Klare (1950) and Converse (1976) showed that no-opinion rates increased as the complexity of the language in a question increased. Converse (1976) also found higher no-opinion rates when questions contained long explanations, when questions required respondents to predict the future rather than simply describing the past or present, for dichotomous questions than for politimous questions (presumably because the former pose more difficulty in terms of describing moderate or qualified opinions), and for questions regarding foreign affairs than for those addressing domestic affairs (presumably because of the greater remoteness of the former and lower knowledge levels likely to be associated with their topics). Furthermore, although Ferber (1966) found no relation between no-opinion rates and position of a question in the questionnaire (in a study that did not manipulate question order experimentally), Culpepper *et al.* (1992) did find high NO rates later in a questionnaire when question order was experimentally manipulated.

Yet another line of empirical inquiry that shares the spirit of the satisficing view was done by Hippler and Schwarz (1989). These investigators proposed that strongly worded no-opinion filters might suggest to respondents that a great deal of knowledge is required to answer an attitude question and thereby discourage respondents who wish to optimize from answering substantively. Hippler and Schwarz (1989) demonstrated that respondents did indeed infer from the presence and strength of a no-opinion filter that follow-up questioning

would be more extensive, would require more knowledge, and would be more difficult. Respondents who said they had no opinion when this option was offered were sometimes also able to report opinions on the same issue later in the questionnaire when the no-opinion option was omitted. Furthermore, these latter responses appeared to be systematic, rather than the result of mental coin flipping.

Still another line of research that calls into question the validity of no-opinion responses is work that has used the proportion of people saying they had no opinion on an issue as an index of the strength of the public's will on the issue. These studies have examined the correspondence between public preferences on a policy issue and government action on that issue, presuming that when the public feels more strongly, government will act more in accord with its wishes. In fact, this expectation has been confirmed using direct questions asking people how strongly they feel about an issue or how important it is to them (e.g., Conover *et al.*, 1982; Schuman and Presser, 1981). But rates of no-opinion responses are unrelated to correspondence between public opinion and government action (Brooks, 1990). This, too, calls into question the notion that no-opinion responses genuinely reflect lack of opinions.

6.5.7 Conclusion

Overall, then, this evidence does not provide a solid basis for recommending that no-opinion filters be included in attitude questions in order to avoid collecting nonattitudes. Rather, it seems that some real attitudes are missed by the inclusion of such filters. Therefore, we view this evidence as supporting the recommendation that such filters be omitted when possible. Of course, if people are to be asked about an obscure issue, on which many are likely to have no information at all, offering a filter may be appropriate and beneficial. But asking people about such obscure matters may not be of much interest to researchers, anyhow.

It is important to note, however, that we believe the essential notion standing behind Converse's (1964) nonattitudes hypothesis is correct: Many people who report attitudes in surveys do not have deeply rooted preferences that shape their thinking and behavior. But we do not think that no-opinion filters are the best ways to identify those individuals. Rather, such individuals are probably best identified by measuring the strength of people's attitudes directly. This can be done in many different ways (see, e.g., Krosnick and Abelson, 1992; Petty and Krosnick, 1995), and the choice of method can be consequential both practically and theoretically. For example, people can be asked how certain they are of their opinions or how important the issues are to them personally or how knowledgeable they are on the topic. We encourage readers to examine the attitude strength literature and to use techniques developed therein to address the nonattitudes problem.

6.6 EPILOGUE

The final word has certainly not been spoken on the design decisions we have considered in this chapter. The number of studies exploring each issue is not great enough to justify final conclusions, and there is occasional disagreement between studies, suggesting that some interacting variables may determine the conditions under which some approaches are better than others. Therefore, there is much work left to be done exploring these issues in future research. We hope that scholars doing questionnaire research will take the opportunity of every data collection effort to incorporate methodological experiments whenever possible. The more we know about differences in measurement approaches, the more effective we all can be in the future in designing our measurement tools.

REFERENCES

Alwin, D. F., and Krosnick, J. A. (1991), "The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes," *Sociological Methods and Research*, 20, pp. 139–181.

Andrews, F. M. (1984), "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach," *Public Opinion Quarterly*, 48, pp. 409–442.

Ayidiya, S. A., and McClendon, M. J. (1990), "Response Effects in Mail Surveys," *Public Opinion Quarterly*, 54, pp. 229–247.

Barrett, R. S., Taylor, E. K., Parker, J. W., and Martens, L. (1958), "Rating Scale Content: I. Scale Information and Supervisory Ratings," *Personnel Psychology*, 11, pp. 333–346.

Bendig, A. W. (1953), "The Reliability of Self-ratings as a Function of Amount of Verbal Anchoring and of the Number of Categories on the Scale," *Journal of Applied Psychology*, 37, pp. 38–41.

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

Bendig, A. W. (1954b), "Transmitted Information and the Length of Rating Scales," *Journal of Experimental Psychology*, 37, pp. 303–308.

Bendig, A. W. (1955), "Rater Reliability and the Heterogeneity of the Scale Anchors," *Journal of Applied Psychology*, 39, pp. 37–39.

Bendig, A. W., and Hughes, J. B. (1953), "Effect of Amount of Verbal Anchoring and Number of Rating-scale Categories Upon Transmitted Information," *Journal of Experimental Psychology*, 46, pp. 87–90.

Bernardin, H. J., LaShells, M. B., Smith, P. C., and Alvares, K. M. (1976), "Behavioral Expectation Scales: Effects of Developmental Procedures and Formats," *Journal of Applied Psychology*, 61, pp. 75–79.

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

Bishop, G. F. (1987), "Experiments with the Middle Response Alternative in Survey Questions," *Public Opinion Quarterly*, 51, pp. 220–232.

Bishop, G. F., Oldendick, R. W., Tuchfarber, A. J., and Bennett, S. E. (1979), "Effects of Opinion Filtering and Opinion Floating: Evidence from a Secondary Analysis," *Political Methodology*, 6, pp. 293–309.

Bishop, G. F., Oldendick, R. W., and Tuchfarber, A. J. (1980a), "Experiments in Filtering Political Opinions," *Political Behavior*, 2, pp. 339–369.

Bishop, G. F., Oldendick, R. W., and Tuchfarber, A. J. (1983), "Effects of Filter Questions in Public Opinion Surveys," *Public Opinion Quarterly*, 47, pp. 528–546.

Bishop, G. F., Oldendick, R. W., Tuchfarber, A. J., and Bennett, S. E. (1980b), "Pseudo-opinions on Public Affairs," *Public Opinion Quarterly*, 44, pp. 198–209.

Bishop, G. F., Tuchfarber, A. J., and Oldendick, R. W. (1986), "Opinions on Fictitious Issues: The Pressure to Answer Survey Questions," *Public Opinion Quarterly*, 50, pp. 240–250.

Brody, C. J. (1986), "Things Are Rarely Black and White: Admitting Gray into the Converse Model of Attitude Stability," *American Journal of Sociology*, 92, pp. 657–677.

Brooks, J. E. (1990), "The Opinion-policy Nexus in Germany," *Public Opinion Quarterly*, 54, pp. 508–529.

Budd, R. J. (1987), "Response Bias and the Theory of Reasoned Action," *Social Cognition*, 5, pp. 95–107.

Budd, R. J., and Spencer, C. P. (1986), "Lay Theories of Behavioural Intention: A Source of Response Bias in the Theory of Reasoned Action?" *British Journal of Social Psychology*, 25, pp. 109–117.

Conover, P. J., Gray, V., and Coombs, S. (1982), "Single-issue Voting: Elite-mass Linkages," *Political Behavior*, 4, pp. 309–331.

Converse, J. M. (1976), "Predicting No Opinion in the Polls," *Public Opinion Quarterly*, 40, pp. 515–530.

Converse, P. E. (1964), "The Nature of Belief Systems in Mass Publics," in D. E. Apter (ed.), *Ideology and Discontent*, New York: Free Press, pp. 206–261.

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

Culpepper, I. J., Smith, W. R., and Krosnick, J. A. (1992), "The Impact of Question Order on Satisficing in Surveys," paper presented at the Midwestern Psychological Association Annual Meeting, Chicago, Illinois.

Dickinson, T. L., and Zellinger, P. M. (1980), "A Comparison of the Behaviorally Anchored Rating and Mixed Standard Scale Formats," *Journal of Applied Psychology*, 65, pp. 147–154.

Dunnette, M. D., Uphoff, W. H., and Aylward, M. (1956), "The Effect of Lack of Information on the Undecided Response in Attitude Surveys," *Journal of Applied Psychology*, 40, pp. 150–153.

Edwards, J. D., and Ostrom, T. M. (1971), "Cognitive Structure of Neutral Attitudes," *Journal of Experimental Social Psychology*, 7, pp. 36–47.

Ehrlich, H. L. (1964), "Instrument Error and the Study of Prejudice," *Social Forces*, 43, 197–206.

Ehrlich, H. L., and Rinehart, J. W. (1965), "A Brief Report on the Methodology of Stereotype Research," *Social Forces*, 43, pp. 564–575.

Feick, L. F. (1989), "Latent Class Analysis of Survey Questions that Include Don't Know Responses," *Public Opinion Quarterly*, 53, pp. 525–547.

Ferber, R. (1966), "Item Nonresponse in a Consumer Survey," *Public Opinion Quarterly*, 30, pp. 399–415.

Finn, R. H. (1972), "Effects of Some Variations in Rating Scale Characteristics on the Means and Reliabilities of Ratings," *Educational and Psychological Measurement*, 32, pp. 255–265.

French-Lazovik, G., and Gibson, C. L. (1984), "Effects of Verbally Labeled Anchor Points on the Distributional Parameters of Rating Measures," *Applied Psychological Measurement*, 8, pp. 49–57.

Garner, W. R. (1960), "Rating Scales, Discriminability, and Information Transmission," *Psychological Review*, 67, pp. 343–352.

Ghiselli, E. E. (1939), "All or None Versus Graded Response Questionnaires," *Journal of Applied Psychology*, 23, pp. 405–413.

Gill, S. (1947), "How Do You Stand Sin?," *Tide* (March 14), pp. 72.

Gilljam, M., and Granberg, D. (1993), "Should We Take Don't Know for an Answer?" *Public Opinion Quarterly*, 57, pp. 348–357.

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

Hartley, E. L. (1946), *Problems in Prejudice*, New York: Kings' Crown Press.

Hippler, H.-J., and Schwarz, N. (1989), "'No-Opinion' Filters: A Cognitive Perspective," *International Journal of Public Opinion Research*, 1, pp. 77–87.

Jacoby, J., and Matell, M. S. (1971), "Three-point Likert Scales Are Good Enough," *Journal of Marketing Research*, 8, pp. 495–500.

Jennings, M. K., and Markus, G. B. (1984), "Partisan Orientations Over the Long Haul: Results from the Three-wave Political Socialization Panel Study," *American Political Science Review*, 78, pp. 1000–1018.

Jennings, M. K., and Niemi, R. G. (1978), "The Persistence of Political Orientations: An Overtime Analysis of Two Generations," *British Journal of Political Science*, 8, pp. 333–363.

Kalton, G., Roberts, J., and Holt, D. (1980), "The Effects of Offering a Middle Response Option with Opinion Questions," *The Statistician*, 29, pp. 65–79.

Kaplan, R. M., Bush, J. W., and Berry, C. C. (1979), "Category Rating Versus Magnitude Estimation for Measuring Levels of Well-Being," *Medical Care*, 17, pp. 501–525.

Kinder, D. R., and Sears, D. O. (1985), "Public Opinion and Political Action," in G. Lindzey, and E. Aronson (eds.), *The Handbook of Social Psychology*, Vol. 2, pp. 659–741.

Klare, G. R. (1950), "Understandability and Indefinite Answers to Public Opinion Questions," *International Journal of Opinion and Attitude Research*, 4, pp. 91–96.

Klopfer, F. J., and Madden, T. M. (1980), "The Middlemost Choice on Attitude Items: Ambivalence, Neutrality, or Uncertainty," *Personality and Social Psychology Bulletin*, 6, pp. 97–101.

Komorita, S. S., and Graham, W. K. (1965), "Number of Scale Points and the Reliability of Scales," *Educational and Psychological Measurement*, 25, pp. 987–995.

Krosnick, J. A. (1991), "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology*, 5, pp. 213–236.

Krosnick, J. A., and Abelson, R. P. (1992), "The Case for Measuring Attitude Strength in Surveys," in J. M. Tanur (ed.), *Questions About Questions: Inquiries into the Cognitive Bases of Surveys*, New York: Russell Sage Foundation, pp. 177–203.

Krosnick, J. A., and Berent, M. K. (1990), "The Impact of Verbal Labeling of Response Alternatives and Branching on Attitude Measurement Reliability in Surveys," paper presented at the American Association for Public Opinion Research Annual Meeting, Lancaster, Pennsylvania.

Krosnick, J. A., and Berent, M. K. (1993), "Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format," *American Journal of Political Science*, 37, pp. 941–964.

Krosnick, J. A., and Fabrigar, L. R. (forthcoming), *Designing Good Questionnaires Effectively*, New York: Oxford University Press.

Krosnick, J. A., and Schuman, H. (1988), "Attitude Intensity, Importance, and Certainty and Susceptibility to Response Effects," *Journal of Personality and Social Psychology*, 54, pp. 940–952.

Lehmann, D. R., and Hulbert, J. (1972), "Are Three-point Scales Always Good Enough?" *Journal of Marketing Research*, 9, pp. 444–446.

Likert, R. (1932), *A Technique for the Measurement of Attitudes*, New York: Columbia University Press.

Lodge, M. (1981), *Magnitude Scaling: Quantitative Measurement of Opinions*, Beverly Hills, CA: Sage Publications.

Madden, J. M., and Bourdon, R. D. (1964), "Effects of Variations in Scale Format on Judgment," *Journal of Applied Psychology*, 48, pp. 147–151.

Martin, W. S. (1973), "The Effects of Scaling on the Correlation Coefficient: A Test of Validity," *Journal of Marketing Research*, 10, pp. 316–318.

Martin, W. S. (1978), "Effects of Scaling on the Correlation Coefficient: Additional Considerations," *Journal of Marketing Research*, 15, pp. 304–308.

Masters, J. R. (1974), "The Relationship Between Number of Response Categories and Reliability of Likert-type Questionnaires," *Journal of Educational Measurement*, 11, pp. 49–53.

Matell, M. S., and Jacoby, J. (1971), "Is There an Optimal Number of Alternatives for Likert Scale Items? Study I: Reliability and Validity," *Educational and Psychological Measurement*, 31, pp. 657–674.

Matell, M. S., and Jacoby, J. (1972), "Is There an Optimal Number of Alternatives for Likert-scale Items? Effects of Testing Time and Scale Properties," *Journal of Applied Psychology*, 56, pp. 506–509.

McClendon, M. J., and Alwin, D. F. (1993), "No-Opinion Filters and Attitude Measurement Reliability," *Sociological Methods and Research*, 21, pp. 438–464.

McCrae, A. W. (1970a), "Information Measures in Perceptual Experiments: Some Pitfalls Encountered by Kintz, Parker, and Boynton," *Perceptual Psychophysics*, 7, pp. 221–222.

McCrae, A. W. (1970b), "Channel Capacity in Absolute Judgment Tasks: An Artifact of Information Bias?," *Psychological Bulletin*, 73, pp. 112–121.

McKelvie, S. J. (1978), "Graphic Rating Scales—How Many Categories?," *British Journal of Psychology*, 69, pp. 185–202.

Miethe, T. D. (1985), "The Validity and Reliability of Value Measurements," *Journal of Psychology*, 119, pp. 441–453.

Myers, J. H., and Warner, W. G. (1968), "Semantic Properties of Selected Evaluation Adjectives," *Journal of Marketing Research*, 5, pp. 409–412.

Narayan, S., and Krosnick, J. A. (1996), "Education Moderates Some Response Effects in Attitude Measurement," Public Opinion Quarterly, 60, pp. 58–88.

Pepper, S. (1981), "Problems in Quantification of Frequency Expressions," in D. Fiske (ed.), *New Directions for Methodology of Social and Behavioral Science: Problems with Language Imprecision*, San Francisco: Jossey-Bass, pp. 25–41.

Peters, D. L., and McCormick, E. J. (1966), "Comparative Reliability of Numerically Anchored Versus Job-task Anchored Rating Scales," *Journal of Applied Psychology*, 50, pp. 92–96.

Peterson, B. L. (1985), *Confidence: Categories and Confusion*, (Report No. 50), Ann Arbor, MI: General Social Survey Project.

Petty, R. E., and Krosnick, J. A. (1995), *Attitude Strength: Antecedents and Consequences*, Hillsdale, NJ: Erlbaum Associates.

Ramsay, J. O. (1973), "The Effect of Number of Categories in Rating Scales on Precision of Estimation of Scale Values," *Psychometrika*, 38, pp. 513–532.

Rosenstone, S. J., Hansen, J. M., and Kinder, D. R. (1986), "Measuring Change in Personal Economic Well-being," *Public Opinion Quarterly*, 50, pp. 176–192.

Schaeffer, N. C. (1991), "Hardly Ever or Constantly? Group Comparisons Using Vague Quantifiers," *Public Opinion Quarterly*, 55, pp. 395–423.

Schuman, H., and Presser, S. (1979), "The Open and Closed Question," *American Sociological Review*, 44, pp. 692–712.

Schuman, H., and Presser, S. (1981), *Questions and Answers in Attitude Surveys*, New York: Academic Press.

Schuman, H., and Scott, J. (1989), "Response Effects Over Time: Two Experiments," *Sociological Methods and Research*, 17, pp. 398–408.

Schwarz, N., and Hippler, H. (1991), "Response Alternatives: The Impact of Their Choice and Presentation Order," in P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: Wiley and Sons.

Schwarz, N., Knauper, B., Hippler, H., Noelle-Neumann, E., and Clark, L. (1991), "Rating Scales: Numeric Values May Change the Meaning of Scale Labels," *Public Opinion Quarterly*, 55, pp. 570–582.

Smith, T. W. (1984), "Nonattitudes: A Review and Evaluation," in C. F. Turner, and E. Martin (eds.), *Surveying Subjective Phenomena* (vol. 2), New York: Russell Sage.

Smith, T. W., and Peterson, B. L. (1985), "The Impact of Number of Response Categories on Inter-item Associations: Experimental and Simulated Results," paper presented at the American Sociological Association Meeting, Washington, DC.

Stember, H., and Hyman, H. (1949–1950), "How Interviewer Effects Operate Through Question Form," *International Journal of Opinion and Attitude*, 3, pp. 493–512.

Stone, D. N., and Schkade, D. A. (1991), "Numeric and Linguistic Information Representation in Multiattribute Choice," *Organizational Behavior and Human Decision Processes*, 49, pp. 42–59.

Taylor, M. C. (1983), "The Black-and-White Model of Attitude Stability: A Latent Class Examination of Opinion and Nonopinion in the American Public," *American Journal of Sociology*, 89, pp. 373–401.

Thurstone, L. L., and Chave, E. J. (1929), *The Measurement of Attitudes*, Chicago: University of Chicago Press.

Wallsten, T. S., Budescu, D. V., Zwick, R., and Kemp, S. (1993), "Preferences and Reasons for Communicating Probabilistic Information in Verbal or Numeric Terms," *Bulletin of the Psychonomic Society*, 31, pp. 135–138.

Watson, D. (1988), "The Vicissitudes of Mood Measurement: Effects of Varying Descriptors, Time Frames, and Response Formats on Measures of Positive and Negative Affect," *Journal of Personality and Social Psychology*, 55, pp. 128–141.

Wedell, D. H., and Parducci, A. (1988), "The Category Effect in Social Judgment: Experimental Ratings of Happiness," *Journal of Personality and Social Psychology*, 55, pp. 341–356.

Wedell, D. H., Parducci, A., and Lane, M. (1990), "Reducing the Dependence of Clinical Judgment on the Immediate Context: Effects of Number of Categories and Type of Anchors," *Journal of Personality and Social Psychology*, 58, pp. 319–329.

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

Wildt, A. R., and Mazis, M. B. (1978), "Determinants of Scale Response: Label Versus Position," *Journal of Marketing Research*, 15, pp. 261–267.

Zaller, J. (1988), "Vague Questions Get Vague Answers: An Experimental Attempt to Reduce Response Instability," unpublished manuscript, University of California at Los Angeles.

CHAPTER 7

Towards a Theory of Self-Administered Questionnaire Design

Cleo R. Jenkins
U.S. Bureau of the Census

Don A. Dillman
Washington State University

7.1 INTRODUCTION

Our understanding of self-administered questionnaire design clearly remains in its infancy. Although recommendations for design have been offered (e.g., U.S. General Accounting Office, 1993; Dillman, 1978), few systematic efforts have been made to derive principles for designing self-administered questionnaires from relevant psychological or sociological theories.

One notable exception is a paper by Wright and Barnard (1975). They reviewed the behavioral research, particularly on language and comprehension, and presented ten rules for designing forms. In a later paper, Wright and Barnard (1978) write that the problems of completing self-administered questionnaires fall into two classes: problems with the language used and problems arising from the way information is arranged spatially. This statement suggests that the spatial arrangement of information is not "language." However, it is more precise to label both as graphic language and to further subdivide them into "verbal" versus "non-verbal" language. One reason for suggesting that the term "graphic non-verbal language" be used is that it is fully encompassing. Not only does it intimate that respondents extract meanings and cues from the

Survey Measurement and Process Quality, Edited by Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, Trewin.
ISBN 0-471-16559-X © 1997 John Wiley & Sons, Inc.