RATING THE RATING SCALES

HERSHEY H. FRIEDMAN, Brooklyn College of the City University of New York
TAIWO AMOO, Brooklyn College of the City University of New York

Rating scales are used quite frequently in research, especially in surveys. Typically, an itemized rating scale asks subjects to choose one response category from several arranged in hierarchical order. Dishonest researchers can, of course, purposefully manipulate the outcome of their research, if they wish, but such biasing may also be totally unintentional. This paper examines issues involved in creating a relatively unbiased rating scale. These include: (1) Connotations of category labels; (2) Response alternative effects; (3) Implicit assumptions of the question; (4) Forced-choice vs. non-forced-choice rating scales; (5) Unbalanced and balanced rating scales; (6) Order effects; (7) Direction of comparison; (8) Optimal number of points; (9) Context effects; (10) Rating approach, e.g., improvement needed, performance, comparison to expectations, comparison to ideal, etc.

INTRODUCTION

Rating scales are used quite frequently in survey research and there are many different kinds of rating scales. A typical rating scale asks subjects to choose one response category from several arranged in hierarchical order. Either each response category is labeled or else only the two endpoints of the scale are "anchored."

Unfortunately, there are many ways that a rating scale can be biased. Dishonest researchers can, of course, manipulate the outcome of their research, if they wish, but such biasing may also be totally unintentional. This paper will describe some of the problems involved in creating a relatively unbiased rating scale.

(1) The Connotations of Category Labels

The words used as descriptors in a rating scale require some thought. Researchers who are interested in creating interval scales (scales in which the respondents perceive equal-sized gradations between the points on the scale) must be careful to choose category descriptors that are truly equal-interval. This is necessary if researchers wish to compute means and use parametric statistics. For example, the following scale is certainly not an equal-interval scale.

__terrible __ horrible __ awful __fair
__slightly good __all right __reasonably good

The perceived psychological distance between "awful" and "fair" is not the same as between "fair" and "slightly good." To overcome this type of problem, researchers have developed lists providing the scale value means of selected adjectives and descriptors that might be used to create rating scales, e.g., Jones and Thurstone (1955), Myers and Warner (1968), Wildt and Mazis (1978). These scales are developed by having subjects rate adjectives on a scale ranging from "the best thing I can say about a product," or "greatest like," to the "the worst thing I can say about a product," or "greatest dislike." One slight problem is that the evaluations are not exactly the same for all groups: Housewives, executives, and students do not all perceive adjectives in precisely the same way. Fortunately, there is a reasonable degree of consistency among the groups (Mittelstaedt 1971).

The Journal of Marketing Management
Volume 9, Issue 3, pages 114-123

These lists containing scale values for descriptors can be very helpful in creating interval scales. For instance, Myers and Warner (1968) found that with housewives, the worst adjectives (the mean on a 21-point scale is in parentheses) were: horrible (1.48), terrible (1.76), awful (1.92), extremely poor (2.08), and exceptionally poor (2.52). The best adjectives were: superior (20.12), fantastic (20.12), tremendous (19.84), superb (19.80), and excellent (19.40). In the middle of the scale: all right (10.76), O.K. (10.28), so-so (10.08), neutral (9.80), and fair (9.52). Schriesheim and Novelli (1989) studied 20 frequency expressions such as: "always," "constantly," "continually," "frequently," ... , "seldom," "not at all," "none of the time," "never." They also computed scale values that would thereby allow researchers to create equal-interval frequency scales.

Bartram and Yielding (1973) tested various general evaluative phrases such as "extremely," "very," "quite," "usually," "fairly," "almost," and "not at all." They found the greatest consistency among subjects when they rated descriptors at the positive extreme rather than the negative extreme, e.g., "extremely" or "totally," rather than "not quite." This result coupled with the tendency for subjects to be more willing to assign positive descriptors, rather than negative descriptors, to stimuli, suggested to the authors that researchers should use descriptors of lesser strength on the negative extreme of the scale. Worcester and Burns (1975) found that descriptors that appear to be grammatical opposites may not be perceived as opposites when presented in a hierarchy of descriptors as part of a rating scale. For example, "tend to disagree" represented a more negative attitude than "tend to agree" represented a positive attitude. Thus, grammatically balanced scales may actually be unbalanced.

Selecting adjectives which distort the scale and its supposed equal psychological intervals might bias research. It would be highly unethical for a researcher to pretest different frequency expressions to find out which one is most likely to "prove" a particular point of view. The expression "seldom" may not produce the same results as say "not often," and "always" may achieve different responses than "continually." In fact, the temporal adverb "occasionally" has been found to be very different than "seldom" but relatively close in meaning to "sometimes" (Schriesheim and Novelli 1989). Similarly, Bradburn and Miles (1979) asked respondents to define what the words "very often," "pretty often," and "not too often" meant with respect to days per month. They found that there was a great deal of variability in how these words are used.

A researcher can influence results by shrewd selection of the endpoints to anchor the scale. For example, Pollack, Friedman and Presby (1990) found that a scale anchored at the endpoints with the adjectives "superior" and "terrible" will not produce the same results as one anchored at the endpoints with the weaker adjectives, "very good" and "very bad." Respondents seem to be reluctant to choose extreme descriptors for their response.

If the rating scales contain numbers, then these numeric values can change the meanings of the scale descriptors. Schwarz et al. (1991) found that the responses of German adults to the question of "How successful would you say you have been in life?" was influenced by the numeric values provided to give meaning to the scale labels. When the scale ranged from 0 ("not at all successful") to 10 ("extremely successful"), 34 percent selected values between 0 and 5. When the scale went from -5 ("not at all successful") to +5 ("extremely successful"), only 13 percent selected the values of -5 to 0. The authors concluded that numeric values ranging from 0 to 10 suggest "the absence or presence of the attribute to which the scale pertains", i.e., degree of success. Negative values from -5 to 0, on the other hand, suggest the presence of the opposite of the attribute, i.e., being a failure. Apparently, even the use of the "right" descriptors for a scale is not enough. Results can be affected by the numeric values attached to the descriptors.

(2) Effect of Response Alternatives on Interpretation of the Question

The response alternatives can affect the interpretation of the question. Several studies have demonstrated that vague questions (e.g., how often respondents were really annoyed), are interpreted differently depending on the frequency of the

response alternatives. Respondents presented with low frequencies of choices (ranging from "less than once a year" to "more than every three months") interpreted "annoyed" as being more severe than those presented high frequency of choices (ranging from "less than twice a week" to "several times a day"). Offering a choice such as "several times a day" indicates to respondents that "annoyed" refers to even trivial cases of annoyance, not only serious episodes (Schwarz et al. 1988; Gaskell, O'Muircheartaigh and Wright 1994).

Knowledge of this phenomenon makes it easy to influence the responses of subjects with questions dealing with relatively ambiguous occurrences, such as "how often have you considered quitting your job?"

(3) Implicit assumptions of the question

Some questions are biased because of an implicit assumption made by the question. Sterngold, Warland and Herrmann (1994) found that the question "How concerned are you about…?" causes a bias in the direction of concern because it assumes that subjects should be concerned about an issue. Using a filter question first asking respondents whether or not they were concerned with an issue and then asking those that were concerned to rate their degree of concern resulted in significantly fewer people showing concern than the former approach.

(4) Forcing a Choice

A forced-choice rating scale will bias results by eliminating the undecideds and/or those with no opinion. Some researchers will purposely leave out the response choice of "undecided," "no opinion," "uncertain," or "don't know." This approach may be reasonable when the researcher has good reason to believe that virtually all subjects have an opinion and you do not want them to "cop out" by indicating they are uncertain. What happens if many subjects are indeed undecided and we do not allow them the option of no opinion? Most will probably select a rating from the middle of the scale, e.g., "average" or "fair." This will cause two biases: (a) it will appear that more subjects have opinions than actually do (b) the mean and median will be shifted toward the middle of the scale. (The "undecided" category is not part of the scale.)

Researchers have found that the public will express their opinions even on "fictitious" issues. Respondents have expressed opinions regarding nonexistent issues such as the Metallic Metals Act (Payne 1951, p.18). Hawkins and Coney (1981) found that respondents would indicate their opinions on various phony topics such as the National Bureau of Consumer Complaints, the proposed Religion Verification Act, etc. They also found that the number of responses to a fictitious issue was affected by the presence of a "don't know" response category. Providing a "don't know" choice significantly reduced the number of meaningless responses.

Tull and Hawkins (1993, p. 379) state that when the researcher believes that respondents truly have no opinion regarding a subject, omitting the "don't know" or "no opinion" category from the scale will provide less accurate responses than if these response options are included. Clearly, studies can be biased by forcing subjects to indicate their opinion when they actually have no opinion and are not simply reluctant to reveal it. It should be noted that political pollsters, when asking respondents who they plan on voting for, virtually always include the "don't know" category. The percentage of subjects selecting this option is a very valuable piece of information. The "undecideds" can swing an election and political analysts are often more interested in the "undecideds" than in people who are already firmly committed to their candidates.

(5) Unbalanced Rating Scales

Generally, rating scales should be balanced, with an equal number of favorable and unfavorable response choices. An example of a commonly used rating scale in business research consists of the following alternatives: "excellent," "very good," "good," "fair," "poor." This scale is unbalanced, with three favorable and only one unfavorable response choice.

Brown, Copeland and Millward (1973) reported that one company used an unbalanced rating scale with the response categories of: "excellent," "extremely

good," "very good," "good," "fair," and "poor." They noted that this scale was "almost guaranteed to provoke an apparently positive response." Needless to say, this scale was not very effective in uncovering the public's opinion regarding products.

One cannot justify an unbalanced rating scale where there is no reason to believe that subjects are just as likely to be negative as positive. The only justification for using an unbalanced rating scale is in a situation where it is known *a priori* that virtually all respondents are leaning in one direction, e.g., brand loyal customers would be expected to be essentially favorable. If you know that one side of the scale will not really be used, you would then want the precision on the side of the scale that will be used. Therefore, an unbalanced rating scale might be called for.

There is some controversy in the literature as to whether people respond solely to the adjective descriptors ("label" effect) or to the relative positions of the response categories to the endpoints ("position" effect). If people's responses are determined solely by the relative position of the descriptors, then the following three rating scales should yield the same means and medians, once the responses are coded 1, 2, 3, 4, 5:

Balanced rating scale: __very good __good __average __poor __very poor
Negatively balanced rating scale: __good __average __poor __very poor __awful
Positively balanced rating scale: __excellent __very good __good __average __poor

There is some research that suggests that subjects are affected by both the label and position effects (Wildt and Mazis 1978). However, a study by Friedman, Wilamowsky and Friedman (1981) which compared the above-mentioned three scales found that the label effect is the stronger effect. In either case, it would be unethical for a researcher to test different types of rating scales until the rating scale that achieves the desired effect is found. Another study by Friedman and Leefer (1981) confirmed that subjects respond much more to the actual descriptor used in the scale rather than to its position relative to the endpoints.

(6) Order Effects in Rating Scales

There is evidence of a bias towards the left side of the scale (Mathews 1929; Holmes 1974; Friedman, Friedman and Gluck 1988). Bipolar rating scales such as warm /cold will not necessarily produce the same results as the same scale in reverse (cold/warm). Traditionally, researchers present the most positive items in the scale first (e.g., "strongly agree," "extremely interesting," or "extremely satisfied") and the most negative items last ("strongly disagree," "very boring," or "extremely dissatisfied"). Belson (1966) examined five different types of rating scales (scales of satisfaction, agreement, interest, approval, and liking), and found that the negative end of a traditional rating scale was used more by respondents when presented first. Friedman, Herskovitz and Pollack (1994) found that a Likert scale with the "strongly agree" response category on the left side resulted in a greater degree of agreement than when the scale was presented to subjects with the "strongly disagree" on the left side. Friedman, Friedman and Gluck (1988) compared three types of semantic differential scales: one with all the favorable descriptors on the left, one with all the favorable descriptors on the right, and one with a random mixing of the scales. The results indicated that placing all the favorable descriptors on the left side of the scale had the effect of shifting responses to the left, that is, toward the more favorable side of the scale.

An unethical researcher interested in manipulating results could place the desired response on the left side of the scale. For example, instead of rating the product as "excellent" to "very poor" the scale would start with "very poor" and end with "excellent." Although we have no way of determining which scale is more valid, we can be reasonably certain that the latter scale will produce more negative evaluations than the former scale.

(7) The Direction of Comparison

Many surveys contain questions of comparison, where respondents are asked to compare two stimuli. For example, subjects might be asked to compare Brand X with Brand Y or Brand Y with

Brand X. It would appear that there should be no difference whether Brand X is compared to Brand Y, in which case Brand X is the subject (the object to be evaluated) and Brand Y is the referent (the object which the subject is told to use as a benchmark), or comparing Brand Y with Brand X (i.e., making Brand Y the subject).

Research by Wanke, Schwarz and Noelle-Neumann (1995) found that the direction of comparison does indeed have a strong effect on response, sometimes reversing the results. For instance, they used two different questions in German on two samples of German students. One question asked: "Thinking of your teachers in high school, would you say that the female teachers were more empathetic with regard to academic and personal problems than the male teachers, or were they less empathetic?" The other group responded to a question with the direction reversed: "Thinking of your teachers in high school, would you say that the male teachers were more empathetic with regard to academic and personal problems than the female teachers, or were they less empathetic?" Responses were measured on a nine-point scale ranging from "less empathetic" (1) to "more empathetic" (9). Not only were the mean ratings statistically different, but when female teachers were the subject, 41 percent of respondents felt that the female teachers were more empathetic than male teachers; when male teachers were the subject, only 9 percent of respondents felt that female teachers were more empathetic than the male teachers. The direction of comparison significantly affected the results obtained when the authors compared soccer with tennis and tennis with soccer on which was the more exciting sport.

The authors concluded that respondents generally "focus on the features that characterize the subject of comparison and make less use of the features that characterize the referent of the comparison." By focusing mainly on the features of the subject, respondents tend to overlook unique features of the referent. Thus, items of comparison that have unique positive features should receive more positive evaluations when they are subjects rather than referents; items that have unique negative features should receive more negative evaluations when they are subjects of the comparison rather than referents.

Later research by Wanke (1996) demonstrated that whether the referent is presented before or after the subject of comparison, there would still be a direction of comparison effect. The direction of comparison order effect is not due to word order.

Although it is not always known to the researcher which features the respondents will focus on, a dishonest researcher might try both directions in a pretest and then know how to slant a study.

(8) The Number of Points

Ideally, a rating scale should consist of enough points to extract the necessary information. Some researchers claim that scales consisting of three points are sufficient (e.g., Jacoby and Mattel 1971). Lehman and Hulbert (1972) claim that when a researcher is interested in averages across people or will combine several individual rating scales in order to create a new scale, then two- or three- point scales are "good enough." If, however, the researcher is working with one rating scale and is interested in individual behavior, more scale points are needed. They recommend the use of a five- to seven- point rating scale.

There is evidence that the more scale points used, the more reliable the scale (Churchill and Peter 1984). Using too few points will result in a scale that is less reliable. However, using more points than subjects can handle will probably result in an increase in variability without a concomitant increase in precision.

Another problem with using too few points in a rating scale is that if two rating scales are then correlated, the observed correlation coefficient will probably be less than its true value. Clearly, information is lost when two continuous variables, X and Y, are treated as discrete variables consisting of only a few points (see Martin 1973, 1978).

How many points should a rating scale consist of? In a literature review, Cox (1980) concluded that there is no single number of points for a rating scale

that is appropriate for all situations. In general, however, he suggested the use of five to nine points. Friedman and Friedman (1986), however, found that in some situations an 11-point scale may produce more valid results than a 3-, 5-, or 7-point scale. Their conclusion was that researchers should consider using anywhere from 5- to 11-point scales.

The answer to the how-many-points question should depend on the stimulus being evaluated. For example, it would be foolish to use more than a three-point scale to rate facial tissues. This is an unimportant, low-involvement product and people think of tissues as being either better than average, average, or worse than average. In rating films, especially with subjects who often go to the movies, an 11-point scale may be appropriate.

A researcher wishing to increase the variability and thereby make it harder for statistics to demonstrate significant differences among stimuli (e.g., comparing different brands of tissues) can accomplish this by using scales with too many points. A two-point scale, on the other hand, used with a stimulus that subjects can actually rate on many gradations will result in a very imprecise measurement. This will make it very difficult to find differences among means. For example, will there be a significant difference between the mean ratings for the presidencies of Abraham Lincoln and William Clinton if the scale consists of only two points, "good" and "bad"?

(9) Context Effects

Many surveys consist of a series of questions whose purpose is to help the researcher determine which factors correlate most strongly with the subjects' overall opinion. Unfortunately, respondents to a survey will often use these previous questions to interpret the meaning of a question and/or to determine what the "proper" answer is supposed to be. For example, if subjects are asked to respond to several questions dealing with the amount of time children waste on moronic television programs, television programs with too much sex, and television programs with excessive violence, and then are asked to rate how they feel about television in general, we would expect them to be more negatively predisposed towards television. These context effects have been shown to bias surveys. In one study, researchers found that subjects, who were first primed by asking them to answer questions dealing with fraud and waste in government programs, were more likely to oppose welfare (Fienberg and Tanur 1989).

Smith (1991) noted that context effects are more likely to occur with questions that: (1) "require wide-ranging memory searches," (2) "access memories that have not been previously organized into a summary evaluation," and (3) "utilize ambiguous terms and/or have certain intent."

One might think that context effects might be eliminated by starting with an overall rating question and then following with negative and/or positive statements which have the ability to prime or influence subjects. However, a study by Schwarz and Hippler (1995) demonstrated that, with a mail survey, responses would be affected whether the "predisposing questions" precede or follow the crucial question. In their study, the crucial question asked German adults the amount of money they would be willing to donate to the "suffering population of Russia this winter." The "predisposing questions" dealt with tax hikes and increased welfare spending. Apparently, self-administered surveys allow subjects to read questions out of order and they may thus be influenced by subsequent questions, not only previous questions. When the same questions were asked in a telephone survey, the "predisposing questions" affected the response regarding donations for Russia only when they preceded this question. Clearly, the use of "predisposing questions" makes it very easy for a researcher to influence the responses of subjects.

Researchers have also studied the effects of order with part-whole questions. This is the case where one item asks about a part of an overall attitude (e.g., marital happiness) and another question deals with the whole (e.g., general happiness). In some studies, an assimilation effect was observed and responses to the later question were found to be more consistent with the responses to the earlier question. Apparently, subjects remembered their

answer to the previous question and based their answer to the later question keeping in mind and emphasizing the earlier response. For instance, asking respondents to first rate their marital happiness and then their general happiness caused subjects to stress their marital satisfaction in the general happiness rating. In other studies, contrast effects occurred, i.e., responses to the later question were pushed in the opposite direction of the previous question. Respondents essentially ignored the earlier question in answering the later question in order to avoid repeating themselves (Tourangeu, Rasinski and Bradburn 1991; Mason, Carlson and Tourangeu 1994).

Mason, Carlson and Tourangeu (1994) found evidence for a contrast effect when a general question about the economy of the respondent's home state followed a similar "part" question about the economy in the respondent's local community. Fewer subjects indicated that they felt that the state economy would get better when the state question followed the community question rather than vice versa. Analysis of responses to an open-ended question indicated that subjects ignored important characteristics of the local economy ("subtraction effect") when rating the state economy. Since respondents were, in general, more optimistic about their local communities than the entire state, this subtraction effect caused a lowering of their evaluations of the state. This is another example of how question order and context might affect the results of the ratings.

(10) Type of Overall Evaluation Question

Typically, there is at least one rating scale that is used to measure respondents' overall feelings towards a stimulus and this is often the single most important question in the questionnaire. This scale can be phrased in numerous ways including:

(1) As an overall performance scale ("Overall how would you rate…") with choices such as "very good," "good," "fair," etc.
(2) As an expectations scale ("Overall, compared with what you expected, how would you rate") with choices such as "much better than expected," "better than expected," "about as expected," etc.
(3) As an improvement scale ("Indicate the amount of improvement, if any, is needed") with choices such as "none," "slight," "some," "much," and "huge."
(4) As a recommend scale ("How likely are you to recommend ____ to a friend") with the response choices being: "very likely," "likely," "neither likely nor unlikely," "unlikely," and "very unlikely."
(5) As a compared to the ideal scale ("Compared to the ideal ___, how would you rate ____?") with the response choices being: "very good," "good," "fair," "poor," and "very poor."
(6) As a "satisfaction" scale ("How satisfied are you with ___?") with the response choices being: "very satisfied," "satisfied," "neither satisfied nor dissatisfied," "dissatisfied," and "very dissatisfied."
(7) As a requirements scale ("How often does using ___ meet your requirements?") with the response choices being: "always meets my requirements," "usually meets my requirements," "occasionally meets my requirements," "rarely meets my requirements," and "never meets my requirements."
(8) As a regret scale ("How often, if at all, do you regret having selected/purchased ____?") with the response choices being: "very often regret," "often regret," "sometimes regret," "rarely regret," and "never regret."

Rust et al. (1994, pp. 61-62) claimed that an expectations question provides more accurate results than the typical performance rating or satisfaction question by greatly reducing the "top box" problem. They felt that one serious problem with performance scales is that oftentimes they are worded so that it is easiest for consumers to choose the top box. However, with an expectations question subjects are less likely to check the top box unless they are truly delighted rather than merely satisfied with the stimulus. Only customers who are truly delighted with a product are likely to use it in the future and/or recommend it to others. This suggests that an expectations scale should produce a mean that is different, and probably lower, than the other scales.

Waddell (1995) suggested that traditional customer satisfaction measurement scales ask the wrong question by focusing on "How am I doing?" rather than "How can I improve?" He claims that consumers usually rate products/services as being better when using performance or satisfaction scales and that these scales often produce high average scores. Neal (1999) posited that satisfaction measures cannot be used to predict loyalty since loyalty is a behavior and satisfaction is an attitude.

Friedman and Friedman (1997) compared the results of six different overall evaluation scales (performance, expectation, recommend, compared to ideal, satisfaction, and regret) and found that, in general, the "compared to ideal" and "expectations" scales result in lower mean evaluations, whereas the "overall performance rating" and "regret" scales produce higher mean evaluations. They also used principal components factor analysis to determine that the six rating scales were indeed measuring the same underlying construct.

Friedman and Rosezweig (1999) compared an overall performance rating scale with an improvement scale. Their results indicated that, in general, the improvement scales produced more negative evaluations than did the performance scales. It seems that subjects were more willing to rate an object as "very good" or "good" than to indicate that the amount of improvement needed by the object was "none" or "slight." Furthermore, respondents were more willing to indicate that the amount of improvement an object needs was "much" or "huge" than to rate the object as "poor" or "very poor." Again, we see that it is not difficult to bias a study by choosing the appropriate overall evaluation scale.

CONCLUSION

There is no doubt that rating scales are used to make very important decisions. We are not talking only about decisions regarding products, but even critical decisions about public policy. It is hoped that researchers will be especially careful to ensure that their scales are as objective and unbiased as possible. At the very least, we hope that this article has sensitized consumers of survey research, such as policy makers and interested individuals, to the difficulties inherent in research based on rating scales.

REFERENCES

Bartram, Peter and David Yielding (1973), "The Development of an Empirical Method of Selecting Phrases Used in Verbal Rating Scales: A Report on a Recent Experiment," *Journal of the Market Research Society*, 15(3), 151-156.

Belson, William A. (1966), "The Effect of Reversing the Presentation Order of Verbal Rating Scales," *Journal of Advertising Research*, 6(4), 30-37.

Brown, Gordon, Tony Copeland and Maurice Willward (1973), "Monadic Testing of New Products-An Old Problem and Some Partial Solutions," *Journal of the Market Research Society*, 15(2), 112-131.

Bradburn, Norman M. and Carrie Miles (1979), "Vague Quantifiers," *Public Opinion Quarterly*, 43(1), 92-101.

Churchill, Gilbert A. Jr. and J. Paul Peter, (1984), "Research Design Effects on the Reliability of Rating Scales: A Meta Analysis," *Journal of Marketing Research*, 21(4), 360-375.

Cox, Eli P., (1980), "The Optimal Number of Response Alternatives for a Scale: A Review" *Journal of Marketing Research*, 17(4), 407-422.

Fienberg, Stephen E. and Judith M. Tanur (1989), "Combining Cognitive and Statistical Approaches to Survey Design," *Science*, 243(2), 1017-1022.

Friedman, Hershey H. and Esther M. Friedman (1997), "A Comparison of Six Overall Evaluation Rating Scales," *Journal of International Marketing and Marketing Research*, 22(3), 129-138.

____ and Linda W. Friedman (1986), "On the Danger of Using Too Few Points in a Rating Scale: A Test of Validity," *Journal of Data Collection*, 26(2), 60-63.

___, Paul J. Herskovitz and Simcha Pollack (1994), "Biasing Effects of Scale-Checking Style in Response to a Likert Scale." *Proceedings of the American Statistical Association Annual Con-ference: Survey Research Methods*, 792-795.

_____ and Joanna R. Leefer (1981), "Label Versus Position in Rating Scales," *Journal of the Academy of Marketing Science*, 9(2), 88-92.

_____, Yonah Wilamowsky and Linda W. Friedman (1981), "A Comparison of Balanced and Unbalanced Rating Scales," *The Mid-Atlantic Journal of Business*, 19(2), 1-7.

_____, Linda W. Friedman and Beth Gluck (1988), "The Effects of Scale-Checking Styles on Responses to a Semantic Differential Scale," *Journal of the Market Research Society*, 30(4), 477-481.

_____ and Seth Rosenzweig (1999), "Biasing Effects in Rating Scales: An Empirical Comparison of Two Overall Rating Scales," *Central Business Review*, 18(2), 20-22.

Gaskell, George D., Colm A. O'Muircheartaigh, and Daniel Wright (1994), "Survey Questions About the Frequency of Vaguely Defined Events: The Effects of Response Alternatives," *Public Opinion Quarterly*, 58, 241-254.

Hawkins, Del I. and Kenneth A. Coney (1981), "Uninformed Response Error in Survey Research," *Journal of Marketing Research*, 18(3), 370-374.

Holmes, Cliff (1974) ,"A Statistical Evaluation of Rating Scales," *Journal of the Market Research Society*, 16(2), 87-107.

Jacoby, Jacob and Michael S. Matell (1971), "Three-Point Likert Scales are Good Enough," *Journal of Marketing Research*, 8(4), 495-500.

Jones, Lyle V. and L. L. Thurstone (1955), "The Psychophysics of Semantics: An Experimental Investigation," *Journal of Applied Psychology*, 39(1), 31-36.

Lehmann, Donald R. and James Hulbert (1972), "Are Three-Point Scales Always Good Enough?" *Journal of Marketing Research*, 9(4), 444-446.

Martin, Warren S. (1973), "The Effects of Scaling on the Correlation Coefficient: A Test of Validity," *Journal of Marketing Research*, 10(3), 316-318.

_____ (1978), "Effects of Scaling on the Correlation Coefficient: Additional Considerations," *Journal of Marketing Research*, 15(2), 304-308.

Mason, Robert, John E. Carlson, and Roger Tourangeau (1994), "Contrast Effects and Subtraction in Part-Whole Questions," *Public Opinion Quarterly*, 58, 569-578.

Mathews, C. O. (1929), "The Effect of the Printed Response Words on an Interest Questionnaire," *Journal of Educational Psychology*, 30, 128-134.

Mittelstaedt, Robert A. (1971), "Semantic Properties of Selected Evaluation Adjectives: Other Evidence," *Journal of Marketing Research*, 8(2), 236-237.

Myers, James H. and W. Gregory Warner (1968), "Semantic Properties of Selected Evaluation Adjectives," *Journal of Marketing Research*, 5(4), 409-412.

Neal, William D. (1999), "Satisfaction is Nice, but Value Drives Loyalty," *Marketing Research*, 11, Spring, pp. 21-23.

Payne, J. D. (1972), "The Effects of Reversing the Order of Verbal Rating Scales in a Postal Survey," *Journal of the Market Research Society*, 14, 30-44.

Payne, Stanley L. (1951), *The Art of Asking Questions*, Princeton, NJ: Princeton University Press.

Pollack, S., H. H. Friedman, and L. Presby (1990), "Two Salient Factors in the Construction of Rating Scales: Strength and Direction of Anchoring Adjectives," *International Conference of Measurement Errors in Surveys*, Tucson, Arizona, November 11-14, 57.

Rust, R. T., A. J. Zahorik, and T. L. Keiningham (1994), *Return on Quality*. Chicago: Probus Publishing Company.

Schriesheim, Chester A. and Luke Novelli, Jr. (1989), "A Comparative Test of the Interval-Scale Properties of Magnitude Estimation and Case III Scaling and Recommendations for Equal-Interval Frequency Response Anchors," *Educational and Psychological Measurement*, 49, 59-74.

Schwarz, Norbert and Hans J. Hippler (1995), "Subsequent Questions May Influence Answers to Preceding Questions in Mail Surveys," *Public Opinion Quarterly*, 59(1), 93-97.

Schwarz, Norbert, Barbel Knauper, Hans J. Hipler, Elisabeth Noelle-Neumann, and Leslie Clark (1991), "Numeric Values May Change the Meaning of Scale Labels," *Public Opinion Quarterly*, 55(4), 570-582.

Schwarz, Norbert, Fritz Strack, Gesine Muller and Brigette Chassein (1988), "The Range of response Alternatives May Determine the Meaning of the Question: Further Evidence on Informative Functions of Response Alternatives," *Social Cognition*, 6, 107-117.

Smith, Tom W. (1991), "Context Effects in the General Social Survey," in Biemer, Paul P. et al. (editors), *Measurement Errors in Surveys*, New York: John Wiley & Sons.

Tourangeau, Roger, K. Rasinski, and N. Bradburn (1991), "Measuring Happiness in Surveys: A Test of the Subtraction Hypothesis," *Public Opinion Quarterly*, 55, 255-266.

Sterngold, Arthur, Rex H. Warland, and Robert Herrmann (1994), "Do Surveys Overstate Public Concerns?" *Public Opinion Quarterly*, 58, 255-263.

Tull, Donald S. and Del I. Hawkins (1993), *Marketing Research: Measurement and Method*, New York: Macmillan Publishing.

Wadell, H. (1995), "Getting a Straight Answer," *Marketing Research*, 7, Summer, pp. 5-8.

Wanke, Michaela, Norbert Schwarz, and Elisabeth Noelle-Neumann (1995), "Asking Comparative Questions: The Impact of the Direction of Comparison," *Public Opinion Quarterly*, 59, 347-372.

Wanke, Michaela (1996), "Comparative Judgments as a Function of the Direction of Comparison Versus Word Order," *Public Opinion Quarterly*, 60, 400-409.

Wildt, Albert R. and Michael B. Mazis (1978), "Determinants of Scale Response: Label Versus Position," *Journal of Marketing Research*, 15(2), 261-267.

Worcester, Robert M. and Timothy R. Burns (1975), "A Statistical Examination of the Relative Precision of Verbal Scales," *Journal of the Market Research Society*, 17(3), 181-197.