THE INFLUENCE OF ADMINISTRATION MODE ON RESPONSES TO NUMERIC RATING SCALES

Roberta L. Sangster, Bureau of Labor Statistics
Todd H Rockwood and Don A. Dillman, Washington State University
Roberta L. Sangster, BLS, 2 Mass. Ave. N E., Rm 4925, Washington DC 20212

Key Words: Rating Scales, Mail Surveys, Telephone Surveys, Mixed Mode Surveys

In a recent study Schwarz and his colleagues (1991a) expanded Dawes and Smith's (1985) seminal work on the properties of scales to examine the influence of numeric rating scale labels on responses. In this study the numeric labels were varied between using all positive numbers (0 to 10) versus a combination of positive and negative numbers (-5 to +5), while the range of numbers were kept equivalent. It demonstrated that people are more likely to choose answers from the higher categories on the scale which includes both negative and positive numbers (-5 to +5) than a scale that used all positive numbers (0 to 10). We extend that work in this paper by introducing a mixed mode comparison between mail and telephone surveys, using 2 new questions and a slightly different range for the scale comparison (-3 to +3 and 1 to 7 as used by Dawes and Smith 1985).

Schwarz et al.'s presented the scale visually on a showcard and read instructions on its use (1991a:572): *How successful have you been in life, so far? Please use this ladder to tell me. This is how it works: 0[-5] means not successful at all and 10 [+5] means that you were extremely successful Which number do you choose?* Sixty-three percent of the respondents responded within the 6-10 range for the 0 to 10 scale, whereas 85% chose numbers within the numerically equivalent 1 to 5 categories for the -5 to +5 scale. As an explanation for this 22% difference, the authors suggest that respondents may use the numeric values to disambiguate the meaning of scale labels. When the zero appears at the end of the 0 to 10 scale, respondents may interpret the zero as the absence of success, while the zero in the center of the scale may indicate the presence of failure (0 to -5 end of the scale) (Schwarz et al. 1991a:572). Thus, while the 1 to 10 scale is interpreted as a unipolar scale, the -5 to +5 scale is viewed as a bipolar scale (Also see Dawes and Smith 1985). The researchers discount the influence of self-enhancement as being a major factor in respondents avoiding assigning themselves a negative number from the bipolar "success" scale, in favor of the explanation that different interpretations are being given to the two types of scales. However, the conclusion about self-enhancement is based on a very small follow-up study that used 22 subjects (Schwarz et. al., 1991a:576).

Hippler and Schwarz (1992:11) subsequently tested the same scales across two modes of administration (mail and telephone), but this time asked a series of questions about 6 politicians: *Please imagine a thermometer that runs from minus five to plus five, with zero in between. Please use this thermometer to tell us how you feel about some politicians. Plus five means that you think very highly of them, and minus five means that you think very little of them. How do you feel about...* The comparison between modes of administration revealed a similar shift toward the higher end of the bipolar scale for both surveys (36% higher for the combined data). Their conclusion was that respondents must interpret the scale's numeric labels regardless of whether presented visually (mail survey) or aurally (telephone survey). A counter hypothesis the researchers discounted was that mail survey respondents would pay more attention to the numeric scale labels because they were presented visually rather than heard aurally as in the telephone survey. If this had been the case they argue, then the telephone would have been less likely to produce the shift in response distributions between the unipolar and the bipolar scales.

O'Muircheartaigh et al. (1993) recently completed research that shows changing the verbal labels about the extent to which the British Advertising Standards Authority should be given more power to control advertisements from a unipolar anchor (not given any more power, given much more power) to a bipolar anchor (given much less power, given much more power) also changes the distribution of responses. The major effect is for fewer people to choose the lower end, and in particular the terminal category when the anchor is shifted to the bipolar scale. Thus, they show that respondents are influenced by the verbal end labels as well as the numerical information provided to them.

Their research also confirms that regardless of the polarity of the verbal labels, or whether respondents are verbally told the endpoints of the scales, that respondents are consistently more likely to choose responses from the 1 to 5 categories (50%, 47%, 55%, and 57% respectively) than the comparable 6 to 10

categories (37%, 34%, 47%, and 50% respectively).[1] This tendency to choose disproportionately from positively labeled (vs. negatively labeled) categories, regardless of verbal anchoring, suggests that other factors may be influencing people's answer choices. Similar to the first study, the respondents in these studies also had the scale presented as a ladder.

Finally, in a much earlier study, Coney (1977) found that the choice for the "best beer" varied by what label was attached to the sample of four beers that were being taste tested. He found that the beer labeled "A" was most favored regardless of where it was positioned in the order of beers (e.g., A, B, C, D, or C, B, A, D ect.). Coney also tested beer's labeled H, L, M, and P. This set of labels did not elicit a label effect, but did vary by position, with the first and third position being more favored.

A possible limitation to Schwarz et. al. (1991) and Hippler and Schwarz (1992) research on the scales is the introduction of visual imagery into the tests (O'Muircheartaigh et al (1993) do not provide the exact wording of the questions). Respondents were asked to use a mental image as part of the test of the scales. Imagining a scale that is a "ladder of success" or a thermometer which suggests "hot" and "cold" may confound any conclusions that can be made from the prior work. The mental image of the thermometer may have changed what is normally an aural processing of information into an aural and visual processing of the scale for the telephone respondents. The theoretical argument for why context effects are more likely to occur in telephone interviews rests on the premise that the telephone interview is an auditory process while the mail survey demands visual processing of information. This argument can also be applied to the research on numeric rating scales.

The theory on context effects (where one response can influence another response) suggests that mail survey respondents will take more time to scan over the survey, review responses, and not be influenced by the presence of an interviewer (See Schwarz, Strack, Hippler, and Bishop 1991b). In contrast the telephone interview relies on auditory rather than visual processing of the survey questions. In the telephone interview the questions are heard sequentially. The respondent does not know what question will be asked next, cannot easily compare responses, and it is more difficult to change answers or reflect on a prior question. In the interview format the interviewer rather then the respondent controls the pace and actual completion of the information. Schwarz et al. (1991b) comment in a different article that "under self-administered questionnaire conditions, ... the respondent is much more dependent on the context that is explicitly provided by the questionnaire to draw inferences about the intended meaning of the questions...and has the time and opportunity to consider related questions to disambiguate the meaning of obscure items. (Pages 6 to 7 draft of chapter). It is logical to think that this process might also occur under the conditions of considering the properties of a scale by the mail survey respondent. If this is true, than the self-administered survey would more likely produce a shift in response distributions than the same questions asked in the telephone survey because the scale properties will experience greater cognitive processing for the mail survey respondent. Hippler and Schwarz (1992) considered this as a possibility in their mode comparison, but discounted this argument when they found similar response distributions between survey modes. However, this conclusion may be somewhat premature given the introduction of the mental imagery of the thermometer in their study. Further testing seems necessary.

Another consideration is the greater potential for socially desirable responses to occur in the telephone and personal interview than the mail survey. Socially desirable responses occur when a respondent chooses a response based on the belief it will place them in a more socially favorable light rather than responding with a more accurate one. Tourangeau and Rasinski (1988:307) discuss this behavior as "editing for the purpose of self-presentation." Schwarz and his colleagues (1991a) attempted to assess this impact, but the results are questionable given the small sample size (n=22, 11 per condition). In contrast to their results, there is a wealth of empirical data that has shown that there are differences in response between the modes of administration, especially when asking socially sensitive or desirable topics (e.g. Schuman and Presser 1981; Aquilino and Loscuito 1990). In general, mail survey respondents are thought to less influenced because of the absence of the interviewer.

Most research on social desirability has focused on what effect administration mode has on the respondent's reporting of personal behaviors (e.g. drinking and driving, drug use, abortion). However, Edwards and Cantor (1991:230) note: *When a survey respondent is reporting the behavior of an establishment with which he or she has some affiliation, a similar tendency may cause him or her to respond so as to present the establishment in a more*

[1] A second set of experimental questions asked about television advertisements being *much more entertaining than the programs (10, +5) or much less entertaining than the programs (0, -5).*

favorable light. Similarly, Israel and Taylor (1990) suggest that ties to a university may have influenced responses. Of interest here is that this study was a mail survey rather than the more vulnerable telephone or personal interview format.

The experimental questions used in these studies not only test the use of negative and positive labels, but also addresses the question whether respondent's will react with socially desirable responses when asked about the college they are attending. We believe that the use of bipolar scales may produce an additive effect for potentially socially desirable questions. This hypothesis suggests that administration mode would likely produce a greater tendency to avoid the negative labeled end of a response scale in the interview where the negative label is only heard than in the self-administered survey format where it is only seen. The counter hypothesis is that the self-administered respondent has more time to think and consider their selection from the rating scale. This should produce greater cognitive processing of the scale properties in the mail survey than the telephone interview which is subject to time pressure and serial processing of the question and response options. Therefore, the mail survey respondent may react more strongly to the variation between a bipolar or unipolar scale. This may produce greater cognitive processing about how the scale is to be used with the question, or it may create a greater reaction to the social implications of choosing from the negative end of the bipolar scale.

Methodology

The data come from two studies conducted by the Social and Economic Sciences Research Center at Washington State University (WSU). Both studies used a split-ballot experimental treatment method and were conducted using two modes of administration to provided a comparison between mail and telephone survey results. Study 1 was conducted in the spring of 1992 using a random sample of all undergraduates at WSU in Pullman, Washington. The sample of 1,200 students was systematically drawn from the Registrar's list of enrolled students. This sample was further subdivided into four subsamples so that 300 students received each treatment. Study 2 was conducted in 1994 in a similar manner, but consisted of a random sample of 700 respondents for the telephone survey and 800 for the mail survey. The sample was drawn from a list of all Seniors enrolled at WSU at that time. The telephone random sample was subdivide into two subsamples of 350 respondents. The subsample for the mail survey was composed of 400 respondents for each treatment. Study 1 dealt with student attitudes towards the University, study habits, and classroom cheating. Study 2 asks about the time needed to complete a degree and internationalizing the curriculum at WSU.

Both studies used the Total Design Method for the mail surveys (Dillman 1978). Mail survey respondents were sent an initial cover letter and questionnaire and a one week reminder postcard. Those who had not responded in three weeks received a replacement questionnaire and cover letter. Response rates calculated as a percent of the initial sample for Study 1 were 65% (n=196) and 60% (n=179) for the mail surveys, and 62% (n=185) and 60% (n=179) for the two telephone surveys. In Study 2, both mail survey's had a 66% response rate (n=262 and n=263) while the telephone surveys yielded a 62% (n=215) and a 60% (n=209) response rate. Both Study 1 and 2 used the following experimental question as the first, lead question:

> In general, how do you rate WSU as a place to get a college education? You can use any number from -3 to +3 to indicate your opinion, with the extremes of -3 meaning VERY UNDESIRABLE and +3 meaning VERY DESIRABLE.

Study 1 asked a second question that read:

> Using the same scale as before, how well do you think the education you are getting at WSU is preparing you for life after you complete college, where -3 means VERY UNPREPARED and +3 means VERY PREPARED.

Study 2 asked a second question that read:

> Using the same scale, how do you feel about the length of time it is taking you to complete a bachelor's degree at WSU, where -3 means MUCH LONGER THAN YOU HAD HOPED FOR and +3 means MUCH SHORTER THAN YOU HAD HOPED FOR.

To create a unipolar scale, the numbers 1 and 7 were substituted for the end points in the other treatments. In Study 1 for the lead question, we added the words, "where zero (or 4) is in the middle" for the telephone respondents. We had thought that adding those words would make the scales more comparable; however, thinking more about the issue of mental imagery we deleted those additional words for the telephone respondents in Study 2.

Results

Our study looks at two issues. The first issue is whether there is a difference in response between the bipolar and unipolar scales. If a difference exists, the second issue is whether responses vary by mode of administration. We first present the results of Study 1, followed by Study 2.

Study 1 Within Mode Comparison

Responses vary between the bipolar and unipolar scale for both experimental questions. Similar to prior

findings, there is greater endorsement at the positive end of the bipolar scale than the unipolar scale. For the lead question there is a striking contrast within the mail survey mode. The last two categories for the bipolar scales shows 71% of the respondents endorsing WSU as a desirable place for an education, while only 50% believe this to be true when using a unipolar scale (Chi. Sq. p. <.02). However, within the telephone mode there is little contrast in responses, 53% for the bipolar and 48% in the unipolar chose the upper end of the scale ranges (Chi. Sq. p.<.98).

Similarly there is a shift in the mail survey for the second question as well. In this instance, there is a 23% shift in responses. Fifty-five percent chose the extreme ranges (2 or 3) to endorse WSU as preparing the student for life when using the bipolar scale, while only 32% used the positive end of the unipolar scale (Chi. Sq. p. <.01). The telephone respondents fall somewhere in between those ranges, with 45% responding with either a 2 or 3 on the bipolar scale and 40% with a 6 or 7 response (Chi. Sq. p <.69).

Study 1 Between Mode Comparison

Responses vary between the telephone and mail survey for the bipolar but not the unipolar scale (Q1 Chi Sq. p.<.21, Q2 Chi Sq. p.<11). The shift in response distributions for the bipolar scale occurs for both experimental questions (Q1 Chi Sq. p.<.02, Q2 Chi Sq. p.<.01). For the lead question, 71% percent of the mail survey respondents chose the last two response categories at the positive end of the bipolar scale compared to 53% of the telephone respondents. There is less of a shift, but still a significant shift of 10% for the second question. The telephone respondents seem to use the bipolar scale similar to the unipolar scale for these two experiments. It should also be noted that in general, telephone and mail survey respondents avoided using the negative end of either scale (unipolar or bipolar).

Study 2 Within and Between Mode Comparison Lead Question

The lead question regarding the desirability of the university for getting an education was repeated in Study 2. The results are somewhat different between the two studies. Again, the mail survey respondents were more likely to use the positive end of the bipolar scale (69%) compared to the unipolar scale (47%) (Chi. Sq. p. <.01). However, in Study 2 this was also true for the telephone respondents (66% bipolar and 50% unipolar, Chi. Sq. p<.01). Hence, there is no difference between the two survey modes in the expected direction. However, the distribution is different for the unipolar scale, with the mail survey choosing numbers from the bottom end of the scale more than the telephone respondents (3% mail surveys versus 8% telephone survey for the last two categories. (Chi. Sq. p.<.01).

Study 2 Within Mode Comparison Second Question

The second experiment in Study 2 asked the respondents (seniors) whether they felt they were taking a longer (-3 or 1) or shorter time period (+3 or 7) to complete their college degree than they had anticipated. Mail survey respondents *did not* tend to use the last two choices for positive end of the bipolar scale (8%) more than those who used the unipolar scale (6%). Counter to prior results, the mail survey respondents tended to use the negative end of the bipolar scale somewhat more often than those who used the unipolar scale (8% difference for the last two categories at the negative end of the scales, (Chi. Sq. p.<.01). In contrast, 26% of the telephone respondents using the bipolar scale chose the last two categories from the positive end compared to 18% using the unipolar scale. While not large, this 8% difference is in the anticipated direction (n.s.).

Study 2 Between Mode Comparison

There is a dramatic difference between modes of administration for the length of time to graduate question. This hold for both scale comparisons. Eight percent of the mail survey respondents chose the positive end of the bipolar scale compared to 26% of the telephone respondents (Chi. Sq. 01). Similarly, 7% of the mail respondents chose the positive end of the unipolar scale compared to 18% of the telephone respondents (Chi. Sq. p. <.01).

Conclusions

Dawes and Smith (1985) concluded that almost any type of rating scale could produce a good estimate of actual height. However, they end their discussion by stating, "It does not, of course, follow that these scales will do a good job of estimating some representational measure of *attitudes* that they purport to evaluate." If nothing else, the present study has illustrated what Dawes and Smith elude to---measuring attitudes with rating scales is a little more complicated than estimating height. Schwarz et. al. (1991a) are probably correct in assuming that the scale labels are subject to different interpretations by some respondents. However, it is less clear when and why this occurs. In Study 1 responses varied with the choice of scale label (unipolar versus bipolar) for both experimental questions for the mail, but not the telephone survey. Study 2 replicated the lead question in Study 1, this time responses varied for both modes of administration, and in almost exactly the same manner (sans the difference in the negative end of the unipolar scale). Finally, for the second question in Study 2 responses varied dramatically between survey modes, but not within survey modes. Clearly, we can

no longer assume that numeric rating scales will necessarily create similar effects across modes of administration. This study has tried to disentangle mental imagery from the scale tests and this may be why the results are different from prior research results. As to why the different results, there are several possible explanations.

It is helpful to apply the principles that Schwarz and others use to explain context effects to comparisons between survey modes and rating scales. The results from Study 1 indicate that the survey respondents may indeed give a little more thought to a rating scale they can see, than one they can't see. However, with greater affiliation and interest to the questions being asked, this may increase cognitive processing and outweigh that lack of visual processing of the scale characteristics. Study 1 largely dealt with cheating behavior and attitudes toward such things as faculty behavior (e.g. helpfulness, availability) or student services (e.g. health services, library). Study 2 was asking Seniors about the problems related to delay of graduation and internationalizing the curriculum. It may be that there was greater interest in this study and that it was viewed as more important and relevant to the respondents. This could explain why the lead question the second time produced a similar result across modes of administration. Respondents in the telephone may have given the scale more consideration due to greater interest and experience with the topic (desirability of the university for getting an education).

We also have to consider that Study 2 may have been more prone to socially desirable responses. This likely explains the dramatic difference in response between the telephone and mail surveys for whether it had taken a longer or shorter time than anticipated to graduate. The telephone respondents, using either scale, said they had taken less time ($\bar{x} = 4.3$) than mail survey respondents ($\bar{x} = 3.5$). It seems the telephone respondents tended to give a "on time" response in contrast to the mail survey respondent who tended to say it took them longer to graduate than expected (likely closer to the truth). This indicates that the telephone respondent may have felt it was important to appear successful to the interviewer by graduating on the time to avoided casting themselves in a less favorable light. Graduating on time would be the socially desirable thing to do.

This study more than ever points out the need to complete mixed mode comparisons that include mail surveys and that use a homogenous population such as students, to help to exclude intervening factors, such as age, education and level of cognitive sophistication. Although, much maligned, a student population can be an excellent choice of study when trying for a homogeneous population. Finally, future research might look at the inclusion and exclusion of visual imagery to understand the influence this has on responses.

References

Aquilino, W. S. and L. Losciuto, 1990. Effects of Interview Mode on Self-Reported Drug Use. *Public Opinion Quarterly,* 54:362:395.

Coney, K. A. 1997. "Order-Bias: The Special Case of Letter Preference." *Public Opinion Quarterly.* 41:385-388.

Dawes, R. M. and T. Smith 1985. "Attitudes and Opinion Measurement." In *Handbook of Social Psychology*, ed. G. Lindsey and E. Aronson. 2:509-566.

Dillman, D. A. 1978. *Mail and Telephone Surveys; the Total Design Method,* John Wiley & Sons, NY. Edwards, W. S. and D. Cantor 1991 "Toward a Response Model in Establishment Surveys." In *Measurement Errors in Surveys"* P. P. Biemer et al. John Wiley & Sons: NY

Hippler, H-J and N. Schwarz 1992. "The Impact of Administration Mode on Response Effects in Surveys." A paper presented at the 47th meeting of the American Association for Public Opinion Research, St. Petersburg, FL.

Israel, G. and C. Taylor 1990. "Can Response Order Bias Evaluations?" *Evaluation and Program Planning*, Vol. 13.

O'Muircheartaigh, C. Gaskell, and D. Wright. 1993. "Weighting Anchors: Verbal and Numeric Labels for Response Scales." Tech. Report No. 6 Methodology Institute, London.

Schuman, H. and S. Presser 1981. *Questions and Answers in Attitude Surveys.* Academic Press, NY

Schwarz, N. et al. 1991a. "Rating Scales: Numeric Values May Change the Meaning of Scale Labels." *Public Opinion Quarterly.* 55:570-582.

Schwarz et al. 199b. "Psychological Sources of Response Effects in Surveys: The Impact of Administration Mode." *Applied Cognitive Psychology* 5:192-212.

Tourangeau, R. and Rasinski 1998. "Cognitive andects Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103: 299-314.

* Opinions expressed in this paper are solely those of the authors.
* Tables are available from the authors.
* More detailed analysis is planned. Study 2 numbers may change slightly, hand-calculated.