THE EFFECTS OF NUMERICAL LABELS ON RESPONSE SCALES

Roberta L. Sangster, Bureau of Labor Statistics
Fern K. Willits, Pennsylvania State University
John Saltiel, Montana State University
Fred O. Lorenz, Iowa State University
Todd H Rockwood, University of Minnesota

Key Words: Numeric rating scales, surveys

The opinion survey, a mainstay of political, sociological, and marketing researchers, has long used rating scales as a means for obtaining subjects' views about various issues (Thurstone, 1927; Likert 1932; Guttman, 1945; Stouffer et. al., 1950; Osgood and Suci 1955, Osgood et. al., 1957; Edwards, 1957). Respondents are asked to indicate their feelings or perceptions on a scale between two opposing descriptors The scale is often presented as a horizontal line, with equally spaced markers between labeled endpoints. While the gradient points may be unlabeled or designated with words, more often they are numbered, using either low-to-high positive integers or minus-to-plus integers with zero in the middle (Figure 1). Use of low-to-high positive coding appears to suggest a simple continuum, while the use of negative-to-positive integers may imply a bipolar conceptualization with zero as the balance point.

Figure 1. Example of Continuum and Bipolar Numeric Rating Scales

Continuum Scale:

Sad 1 2 3 4 5 6 7 Happy

Bipolar Scale:

Sad -3 -2 -1 0 1 2 3 Happy

Occasionally, graphical representations (e,g,, smile faces, ladders, thermometers) are employed to aid in the interpretation of the scale.

What are the implications of these different numerical labels and formats for the answers given by respondents? Using our example, do people indicate different happiness levels depending upon whether the item is presented in a bipolar format or as a continuum? Considerable debate has arisen about this issue, but only a few studies have addressed the question empirically. The current analysis provides additional data drawn from recent research concerning the effects of various numerical labels on rating scales in self-administered surveys.

Previous Research

Schwarz and his colleagues (1991a) described two experiments in which, using a split ballot design, German subjects were offered response categories presented as either eleven-point continua (coded 0 to 10) or bipolar (-5 to +5) scales in personal interviews. A showcard was handed to the respondent and instructions were read about its use (Schwartz et. al., 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 answered within the 6 to 10 range for the continuum scale, whereas 85% chose numbers within the locationally equivalent 0 to 5 categories for the bipolar scale. The researchers suggested that people used the numeric properties of the scale to interpret the meaning of the question. When the zero appeared at the low end of the 0 to 10 continuum scale, respondents may have interpreted zero simply as the absence of success, while in the bipolar scale, the low score (-5) may have been seen as not simply the absence of success, but the presence of failure. The second experiment used a self-administered questionnaire, a sample of German university students, and similar scales to ask about the success and childhood happiness of the subjects and their perceptions of their parents' success and childhood happiness. To check the effects of differing endpoints, the "low" end of the scale was, in some instances, labeled *Unhappy* or *Unsuccessful*; in other cases *Not so happy* or *Not so successful* were used. Again, they found that subjects were less likely to choose responses in the middle or lower end of the bipolar scales than they were when the scale points ranged from 0 to 10.

O'Muircheartaigh et al. (1993) completed a study about the amount of power that should be given to the British

Advertising Standards Authority to control advertisements. The study was designed to examine how word anchors might signal to the respondent whether the scale was "unidirectional or bi-directional." They felt a bipolar anchor combined with a bipolar scale and continuum anchor with a continuum scale would work better for respondents than a mis-match of word anchor and scale labels (i.e., bipolar anchor with continuum scale and continuum anchor with bipolar scale). The question used in the study asked 2165 respondents... *to what extent do you think the Advertising Standards Authority should be given more power to control advertisements*. Again the scales were illustrated on a showcard as a ladder. The experiment used the same word anchor at the positive end for both scales (*given much more power*). Either the bipolar word anchor, *given much less power*, or the continuum word anchor, *not given any more power*, were assigned to the lower end of the scales. This created a four-way comparison (2x2 design). They anticipated an interaction effect for the mixed set comparison, but this was not the case. Instead they found the bipolar word anchor increased the percentage (about 10%) at the midpoint for both scale types (continuum and bipolar scales). The bipolar word anchors appeared to decrease the percent at the lowest endpoint (0,-5)by about six percent. Other than these two differences, the four distributions for the scales were nearly identical. (O'Muircheartaigh et al. 1993:12).

Schwarz, et al. (1991b)) tested the two types of numeric rating scales across two modes of administration (mail and telephone surveys). This study asked a series of questions about six 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 survey modes (36% higher for the combined data). It was unclear whether the mail survey showed the scale as a thermometer, or whether respondents were merely asked to imagine a thermometer.

In a cognitive research study, Stinson (1998) tested the thermometer and the ladder along with eight other visual scale graphics (i.e., faces scale, the delighted/terrible scale, circles scale, worry scale, positive and negative line, and the pie scale).[1] Forty respondents answered a series of 14 economic well-being questions.[2] Stinson debriefed the respondents who used the thermometer scale and found that sixty-percent of the participants had a positive reaction to the thermometer scale.[3] Respondents said that the temperature labels were clear and easy to follow and that they had no difficulty selecting an appropriate response. The other forty-percent had moderately strong negative reactions. Respondents were confused by the references to temperature that evoked images of climate temperatures and unsatisfied with the mid-point of the scale. She concluded that respondents tended to be more divided in how they used the thermometer compared to the other scales she tested. Respondent's reaction to the ladder scale (9-point) was also mixed (negative and positive reactions). Stinson summarized these results by stating, "test participants appear capable of using the Ladder Scale effectively and providing temporal comparisons of their financial situations. However, considering the strong negative reactions to the scale and questions, one is led to question the actual value of this approach" (1998: 28).

These previous studies suggest that the use of differing numeric scale labels on rating scales may impact on the types of answers that are obtained, with bipolar scales tending to yield a more skewed distribution then do continuum rating scales; word labels may also influence response distributions. However these findings were based on only a few studies, some of which involved personal interviews rather than self-administered surveys. Moreover, some of the findings may have been confounded by the addition of the visual scales. Stinson's research clearly indicated that visual scales, such as thermometers and ladders, influenced how numeric scales were interpreted. Additional data are needed to further explore the issue and to test the limits of generalizeability.

The purpose of the current analysis was to add to the body of knowledge concerning the effect of differing numerical labels on responses to mail surveys by pre-

[1] The positive-negative line did not include integers food, cost of transportation, cost of health care, cost.

[2] How do you feel about the…cost of shelter, cost of of clothing, cost of utilities, cost of recreation. How do you feel about your…total family income, savings, investments, financial security, financial situation taken as a whole, financial future, chances of getting ahead financially.

[3] The use of the "feeling thermometer" comes from the National Election Studies, which has used the scale since the 1950s.

senting the results of a number of experiments that replicated and extended the findings of previous research. In each of these experiments, the numerical rating scales were presented without the addition of visual aids. For consistency, the studies all used seven-point scales. However, there was variation in the questions used and in the populations studied with the goal of examining the issues across a variety of situations. One study provided data for extending the study by O'Muircheartaigh et al. by adding word labels to each scale point.

The Studies [4]

A total of eight studies, containing 24 experiments were carried out. Five studies (9 experiments) used mail or self-administered surveys and samples of university students in Iowa (n=703), Pennsylvania (n=1052, n=1071) or Washington (n=375, n=517). Two studies (4 experiments) involved mail surveys of university faculty members (n=1084, n=1016) at various campuses in Pennsylvania. One study (11 experiments) was a mail survey of Montana farmers and ranchers (n=1022).

For each of the experiments sample members were randomly assigned to one of two treatment categories. Half of the subjects were asked to respond to one or more questions using a bipolar rating scale (-3 to +3); half were given a continuum rating scale in which the gradients were numbered from 1 to 7. For four experiments in the Montana farm and ranch survey, in addition to the numerical scores, word designations were included.

Topics included student and faculty evaluations of the "desirability" of their university as a place to get an education, the extent to which they believed that this education "prepared" students for life after college, and the length of time taken to complete the degree relative to their expectations. The Montana farm and ranch study asked about how "harmful" or "beneficial" certain changes in agriculture would be to farmers and ranchers.

Results

University Student and Faculty Surveys

Seven surveys (12 experiments) used samples of university students and faculty. Subjects were requested to rate their universities as a place to get an education on a scale from *Very Undesirable* (-3 or 1) to *Very Desirable* (+3 or 7) (Table 1). In every case, the proportion of responses above the mid-value on the scale was greater for the bipolar format (Bipo>Mid) than for the continuum (Cont>Mid). In contrast, the continuum format was associated with proportionally greater use of the midpoint value (Cont Mid) in comparison to the midpoint value of the bipolar scale (Bipo Mid) and, in most instances, greater use of points below the midpoint (Cont<Mid and Bipo<Mid). For 5 of the 7 experiments, these differences between the two rating scale formats were statistically significant ($p<.05$); an additional one nearly reached significance ($p=.062$). Overall, students and faculty tended disproportionately to avoid the negative scale numbers and, as a result, rated the university higher when using the bipolar format than when the continuum was presented.

Table 1: Desirability of University Question[5]

Sample	Cont. <Mid	Bipo. <Mid	Cont. Mid	Bipo. Mid	Cont. >Mid	Bipo. >Mid
	---------- % ----------					
*Students	5	4	11	4	84	92
#Students	6	3	10	8	84	89
* Students	8	6	9	5	83	89
*Faculty	6	4	12	7	82	89
Senior s	7	6	12	8	81	86
*Students	15	10	10	6	75	84
*Faculty	10	10	21	13	69	77

* χ^2 $p < .05$

\# χ^2 $p = .062$

A second question on five of the student and faculty surveys asked subjects to rate how well they felt their universities were preparing them for life after college. The end-points of the scale were: *Very Unprepared* (-3 or 1) and *Very Prepared* (+3 or 7). Again, both students and faculty were less likely to choose ratings below the mid-values when using the bipolar scale in comparison to the continuum scale indicating their relatively greater reluctance to answer in terms of zero or negative codes (Table 2).

[4] Western Regional Project W-183, "Improvement of Rural and Agricultural Sample Survey Methods." This is a multi-state consortium of faculty from land grant university agricultural experiment stations and others working together to conduct replicative experimental research on measurement error in survey. See Lorenz and Bruton (1996); Sangster, et al. (1994);Willits, et al. (1998).

[5] < Mid = below midpoint
Mid = midpoint
> Mid = above the midpoint

Table 2: University Preparation for Life Question

Sample	Cont. <Mid	Bipo. <Mid	Cont. Mid	Bipo. Mid	Cont. >Mid	Bipo. >Mid
	---------- % ----------					
*Faculty	7	5	15	8	78	87
*Student	7	9	16	5	77	86
*Student	8	6	17	11	75	83
*Student	11	9	16	11	73	80
*Faculty	11	12	24	16	65	72

* χ^2 p <.05

In one study, Washington State asked college seniors how they felt about the length of time it was taking them to complete their bachelor's degree. This question was worded in such a way that the more likely response would occur for the lower integers for both scales (i.e., -3 or 1) meaning that is was taking *Much Longer* then anticipated; the high end of the scale (+3 or 7) meant that it was taking a *Much Shorter* time then anticipated. Seniors were more likely to say that it was taking longer to graduate when using the bipolar scale (51%) then when using the unipolar scale (39%). In this case the direction of the effect was toward the greater endorsement of the negative end of the bipolar scale (Table 3). This suggests that the attenuation of bipolar scales can occur for the negative values as well. While the difference was not large, it could affect the substantive conclusion of the study. This is a troublesome finding for survey practitioner, because it suggests that subject tendency to avoid zeroes and negative responses may apply only to items about which one would tend to hold overall positive views.

Table 3: Seniors Length of Time to Graduate Question

Scale Range	Continuum	Bipolar
	---------- % ----------	
<Mid	20	15
Mid	41	34
>Mid	39	51

* χ^2 p <.05

Montana Farmers and Ranchers

A mail survey of Montana farmers and ranchers contained two sets of questions asking how various changes in the Montana cattle industry would be expected to affect them. End points on the rating scales were *Very Harmful* (-3 or 1) and *Very Beneficial* (+3 or 7). The first set asked about policy issues that would largely affect ranchers (11 items). The second set of experiments (4 questions) asked about issues of concern to both farmers and ranchers. This latter set of experiments also had word labels assigned to each integer.

For the first set of questions, when the entire sample was used, six of the seven experiments yielded differences in response distributions (Table 4). There was a tendency for respondents to disproportionately avoid the low (very harmful) end of the scale when it was labeled with negative numbers. However, there was also a greater use of the midpoint using the bipolar scale, and little difference between the two formats in the tendency of subjects to select integers above the midpoint

Table 4: Harmful-Beneficial Questions Farmers & Ranchers

	Cont. <Mid	Bipo. <Mid	Cont. Mid	Bipo. Mid	Cont. >Mid	Bipo. >Mid
	---------- % ----------					
*Export	2	1	10	15	89	84
*Feedlot	5	2	17	23	78	75
*Market Ed.	10	4	26	31	64	65
*Bkg Abil.	6	2	29	33	65	65
*Cows	9	4	30	36	61	60
*Value Ad.	10	5	36	41	54	54
Video	12	9	46	51	42	40

* χ^2 p <.05

Because these questions referred specifically to issues confronting cattleman, the data were re-run using only ranchers (eliminating crop farmers). When this was done, a pattern consistent with the experiments presented in Tables 1 and 2 was found, although only two of the experiments were significant (Table 5). Since the questions had greater relevance to ranchers then farmers, salience might have been an intervening factor to consider as an explanation for these results.

Table 5: Harmful-Beneficial Questions Ranchers Only

	Cont. <Mid	Bipo. <Mid	Cont. Mid	Bipo. Mid	Cont. >Mid	Bipo. >Mid
	---------- % ----------					
Export	2	1	7	6	91	93
Feedlot	4	1	16	16	80	83
*Market Ed.	9	4	26	22	65	74
Bkg Abil.	6	3	27	28	67	70
*Cows	11	5	29	27	60	68
Value	12	6	33	34	55	60
Video	11	10	40	39	49	51

* χ^2 p < .05

Crop farmers (for whom the questions were not applicable) were more likely to choose the midpoints coded zero and their answers were responsible for the overall pattern observed in the total sample. Apparently "zero"

represented a clearer "no effect" response on the bipolar scale than did "4" on the 1 through 7 scale.

For the four experiments where all points on the scale were given word labels in addition to the numerical codes, none of the four items presented significant format effects (Table 6). This suggests that the addition of words to the numerical scales nullified the effects of differing scale values.

Table 6: Harmful-Beneficial Questions with Word Labels

	Cont. <Mid	Bipo. <Mid	Cont. Mid	Bipo. Mid	Cont. >Mid	Bipo. >Mid
	---------- % ----------					
World Mk	28	33	10	9	61	58
Farm Bill	30	34	31	24	39	42
Flex Acres	44	42	43	42	13	16
Grazing	49	51	42	37	9	12

Chi. Sq. n.s. findings

Conclusions

Most of the findings reported here are in accord with previous research concerning the tendency of subjects to disproportionately avoid the mid-value (zero) and negative end of a bipolar scale. As a result, evaluations of the three universities by their students and faculty were more positive when using the bipolar scale. Moreover, data from the Montana study suggested that this tendency might apply not only to self/significant evaluations, but to other descriptive ratings as well. This was consistent with the findings of O'Muircheartaigh and his colleagues study of advertisements.

However, several caveats to this generalization seem warranted. First, virtually every experiment reported here (and in previous studies) dealt with distributions in which most subjects reported scores above the mid-value on the scale. In the single instance in which the distribution was skewed in the opposite direction, the pattern of avoiding negative and zero values on the bipolar scale did not hold, and indeed was reversed. While a single instance of reversal does not establish a pattern, it is noteworthy. Second, the salience of an item may impact on the scale format effect. For questions that are of little or no relevance, subjects may be more rather than less likely to utilize the mid-value of zero on a bipolar scale than to choose a positive integer on a continuum format to represent "no effect".

Clearly additional research on these issues is needed to further understand the nature and meaning of these scale format effects and to explore the types of situations in which they are evidenced. To what extent are the current findings relevant to other types of items such as beliefs about the efficacy of different programs, estimates of priorities to be given to various alternatives, or the extent to which subjects agree or disagree with selected issues? Are respondents to *telephone* surveys similarly influenced by differing numeric codes on rating scales? What are the cognitive processes involved in subject reluctance to choose responses that are designated by zero or negative numbers. To what degree do differences in scale formats affect the *relationships* of the measured variables to other factors, both independent and dependent variables.

Recommendations to Researchers

Given the findings that responses to rating scales are affected by the numeric labels used to designate the gradients along the continuum between two named endpoints, what should researchers do?

Should researchers avoid the use of these types of rating scales and label all response categories with words?

There is nothing in the research that would support the abandonment of these types of rating scales. Use of a graduated scale with equidistant markings suggests the idea of equal intervals for the resulting scale more clearly than would be possible with any word responses. Moreover, using numeric responses means that the data are precoded, thus simplifying and reducing errors in data preparation.

What system of numeric coding should be used?

It seems reasonable to use a numbering format that is in accord with the desired nature of the scale. Negative-to-positive coding implies a bipolar concept and hence these labels are most appropriate when the endpoints clearly designate opposites and the mid-value of zero (0) has meaning. . This would be true, for example, if the concept being measured dealt with issues that refer to both "profit" or "gain" and "loss." If, however, the concept being measured is a continuum with this low end of the rating scale representing the absence of the attribute, while the high end stands for "a great deal," positive numbers from low-to-high better represent the concept being measured.

What about visual graphics?

Most of the studies presented here produced results similar to those found in previous research, but the magnitude of the differences appeared to be smaller, and did not always reach statistical significance. Perhaps the use of visual images influenced the responses in prior studies, enhancing the observed format distinctions.

Are there times when a bipolar concept would be better assessed using a rating scale with low-to-high positive integers rather than one with negative-to-positive scores?

The reluctance of subjects to select negative codes can mean that part of a bipolar (negative-to-positive) scale will be virtually unused. In such cases, the spread of the scale values may be attenuated, leading to a relatively high mean score and a reduced variance. If it is anticipated that very few subjects will choose the negative scores, it may be more useful to utilize a single continuum scale with endpoints that deal only with the presence or absence of the positive characteristic. Moreover, even in instances which appear to be bipolar (e.g. sad/happy), it may be useful to treat the endpoint descriptions as separate dimensions rather than extremes of the same continuum. Thus, being "happy" does not necessarily mean the absence of "sad." Using two separate continua, one asking for "happiness rating" and one for a "sadness rating" might improve both the measurement of these ideas and contribute to their conceptualization as well

References

Edwards, A. L., 1957, *Techniques of Attitude Construction,* Appleton-Century and Croft, Inc. NY.

Guttman, L., 1945, "The Basis for Scalogram Analysis," In *Measurement and Prediction: Studies I Social; Psychology in World War II, Princeton: Princeton University Press,*4:60-90.

Likert, R. A., 1932, "A Technique for the Measurement of Attitudes," *Archives of Psychology,* No. 140.

Lorenz, F.O. and Bruton, B. T., 1996. "Using Experiments Within Mass Class Surveys in Teaching Research Methods." *Teaching Sociology 24:3.*___

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

Osgood, C. and Suci, F., 1955, :Factor Analysis of Meaning, *Journal of Experimental Psychology,* 50:325-338.

Osgood, C. , Tannenbaum, P., and Suci, G. 1957, *The Measurement of Meaning,* Urbana, IL: University of Illinois Press.

Sangster, R. L., Rockwood, T. H, and D. A. Dillman, 1994. "The Influence of Administration Mode on Responses to Numeric Rating Scales," *American Statistical Association Proceedings of the Survey Methods Sections*, pp. .

Schwarz, N., B. Knauper, H. J. Hippler, E. Noelle-Neumann, and Clark, L., 1991a, "Rating Scales: Numeric Values May Change the Meaning of Scale Labels," *Public Opinion Quarterly 55(4):570-582.*

Schwarz, N., Strack, F, Hippler, H-.J., and Bishop, G. 1991b., "Psychological Sources of Response Effects in Surveys: The Impact of Administration Mode," *Applied Cognitive Psychology* 5:192-212.

Stinson, L. L. 1998, *Subjective Assessment Of Economic Well-Being: Wave II,* Technical Report for the Bureau of Labor Statistics, Washington, DC.

Thurstone, L. L. 1927. "A Law of Comparative Judgement," *Psychological Review,* 54, No. 3.

Willits, F.K., Sangster, R. L. and Saltiel, J. 1998, "Coding and Meaning: Positive and Negative Scale Labels." Presented at the Annual Meeting of the Rural Sociological Society, Portland, Oregon.