

EFFECTS OF PRECODING RESPONSE OPTIONS FOR FIVE POINT SATISFACTION SCALES IN WEB SURVEYS

Mario Callegaro, Tom Wells and Yelena Kruse
Knowledge Networks, Inc., Menlo Park, CA

ABSTRACT

Previous research shows that changes in number-label association can affect respondents' answers. In our first study, conducted with members of KnowledgePanel®, we manipulated the order of presentation of a fully labeled five point satisfaction scale (very satisfied to very unsatisfied); the association of numbers from 1 to 5 to each scale point; and we also reversed the number order (from 5 to 1). This manipulation created six experimental conditions. For each condition, we presented seven different items on separate screens. The completion rate for this study, done in April 2008, was 71%.

In the second study we used only a polar point satisfaction scale and manipulated the order of the scale as well the number-label associations (1 to 5 or 5 to 1). This created 4 experimental conditions to which KnowledgePanel members were randomly assigned during a survey conducted in August of 2008 with a completion rate of 72%. For each condition, we presented seven different items on separate screens.

In the first study we investigated the differences between each version, focusing specifically on the following comparisons: response options with numbers associated vs. options without numbers; order of the response option presentation; and order of the number association. Results from experiment 1 suggest that (1) reversing the order of presentation makes a significant difference resulting in primacy effects and in an increase of time latency when the scale is shown starting from "Very unsatisfied". (2) Changing the number-label association does not make much of a difference; it appears that for a fully labeled scale the label is almost overriding the number. (3) Respondents are, however, slightly confused when the number-label association does not go in the expected direction.

In the second study we compared the 4 experimental groups, first between (same order of presentation but reversing the number-label association), and then within (same number-label association but reversing the order of presentation of the scale). We did not find many significant differences among groups, thus leading us to conclude that for a polar point scale the number-label association has less of an impact than for a fully labeled scale.

The results are discussed in light of current theories of question answering processes and visual layout.

INTRODUCTION AND LITERATURE REVIEW¹

In self-administered paper questionnaires it is common practice to number the response options, a practice also called precoding. For example, Bourque and Fielder (1995) suggest “do not avoid precoded response categories, but clearly indicate the code that corresponds to each response” (p. 102). Dillman (2007), in delineating principles for constructing questionnaire pages for self administered paper instruments, has a specific principle on this topic: “Principle 3.21: use numbers or simple answer boxes for recoding the answer” (p. 124). The reason for this advice is to simplify traditional data entry, although it is not really necessary if the questionnaires are optically scanned.

In web surveys there is really no need to precode response options, but it is still a common practice, and any reader with limited experience in answering web questionnaires can name some examples. The goal of this paper is to shed some further light on the relationship between the numbers used to precode response options, and the meaning of each response alternative. Our study finds its place in the literature of context effect (Smyth, Dillman, & Christian, 2007 for a review), and more specifically, on the numeric values of response alternatives and order effects within rating scales. We manipulated both number association and order of presentation for items with a five point satisfaction scale.

In one of the first studies we found on the relationship between precoded numbers and answer scale labels, Abrams (1966) manipulated an 11 point scale with polar point (also called endpoint) labels. In one condition respondents were mailed a questionnaire asking them to rate six products where the scale was containing numbers from -5 to +5. The scale was labeled with endpoints (definitely dislike and definitely like), plus the middle label (neutral). In another condition the numbers went from 1 to 10 with the same endpoint labels, but this time with no neutral midpoint label. Although the middle label confounds the comparability of the two conditions, the mean of the first experiment across the items was 2.3, while in the second case, it was 7.4.

Schwarz and colleagues (1991) changed the numbers associated to a polar point labeled scale, asking a sample of respondents in a face-to-face surveys how successful they have been in life. The question was reproduced on a showcard with endpoints labeled “not at all successful” and “extremely successful”. In one condition the numbers associated ranged from 0 to 10, while in the other condition the numbers ranged from -5 to +5. In the first case 34 percent of respondents chose values from 0 to 5, while in the second condition, only 13 percent endorsed values formally equivalent between 0 to 5 (in this case -5 to 0). The authors conclude that respondents use the meaning of the numbers to disambiguate the meaning of answer scales, and to help interpret the question. In a later study the same effects were found for both mail and telephone interviews, thus leading the authors to conclude that numeric values do not need to be presented in a visual format to receive sufficient attention (Schwarz & Hippler, 1995). Schaeffer and Baker (1995) were able to obtain similar findings in a telephone interview, using either a 1 to 7 scale or a -3 to +3 scale. The authors also introduced a label for the middle category (4 or 0: “neither oppose nor favor”) and found that that reduced the shift of the distribution to the right side (with the -3 ... + 3 scale).

¹ This paper is an extension of previous work presented at the Annual Midwest Association for Public Opinion Research (MAPOR) conference in Chicago on November 21–22nd 2008

In an unpublished experiment cited by O'Muircheartaigh, Gaskell and Wright (1995), the authors compared a [0...6] vs. [1...7] and a scale [0...10] vs. [1...11], noticing a weaker effect for the first in comparison to the second experiment. The same authors (O'Muircheartaigh et al., 1995) confirmed a previous study by Schwarz et al. (1991), with a face-to-face interview where the question was presented on a showcard as a vertical ladder². The new part of the study [experiment one] was to signal explicitly the numeric association in the question wording [the scale ranges from 10 (-5)...] or not. Explicit signaling was not found to produce a higher shift in the distribution, therefore providing some evidence that the respondents use the numbers associated with the label to clarify the meaning of the question, even if they are not instructed to do so. In experiment two, the authors manipulated the endpoint of the scale from unipolar to bipolar leaving everything else constant³. The effect of the bipolar scale was to increase the percentage at the midpoint of the scale at the expense of the lower end of the scale. Moreover the distribution with the highest mean was when the bipolar labels were associated with -5 and +5.

Schwarz and colleagues (1998) wanted to explore how respondents interpret the meaning of vague quantifiers such as "rarely" and "often" when associated with numbers, either going from 0 to 10 or from 1 to 11. A sample of undergraduate students from the University of Michigan was randomized to the two conditions answering three low behavioral frequency questions: How often do you (1) get a haircut? (2) Visit a museum? (3) Attend a poetry reading? The students did indeed interpret the number value of "rarely" as lower frequency when associated with 0 than when associated with 1.

Tourangeau, Couper and Conrad (2007) replicated the Schwarz experiment in a web survey, also adding colors to the labels. In one condition a seven point scale had a color scheme of shades of a single hue (from dark blue to red blue), while in another condition it was three shades of red for the left side of the scale, white shade for the middle point, and three shades of blue for the right part of the scale. The darker shades were the extremes of the scale. The authors provided additional evidence of the impact of the negative numerical labels, no matter if the scale was clearly unipolar, clearly bipolar, unclear to polarity, or with different hue shades.

Christian (2003; Dillman & Christian, 2005) manipulated a 5 point desirability and satisfaction scale in a web survey of Washington State University (WSU) students, going from 1 ("Very Desirable" or "Very Satisfied"), to 5 ("Very Undesirable" or "Very Dissatisfied"). In one condition the scale was presented fully labeled with numbers⁴, in a second condition the numbers were kept, but the labels were only displayed for the endpoint of the scale. In the third condition only the polar point labels of the scale were shown without numbers. The first condition produced significantly higher positive scores in comparison to the polar point labels, with or without numbers. The authors also did not find any difference between the last two conditions, i.e. polar point with number vs. polar point without numbers. All items were presented as one question per screen and with response categories listed vertically. In another experiment with WSU students, Christian, Dillman & Smith (2008) this time reversed the order of the number association

² How entertaining do you think the adverts on television are, compared to other programs? (Much more entertaining than the programmes) vs. (Much less entertaining than the programmes).

³ ... to what extent do you think the Advertising standards Authority should be given more power to control advertisement? (not given any more power, given much power) vs. (given much less power, given much more power).

⁴ [1 Very Desirable; 2 Somewhat Desirable; 3 Neither Desirable nor Undesirable; 4 Somewhat Undesirable; 5 Very Undesirable] [1 Very Satisfied; 2 Somewhat Satisfied; 3 Neither Satisfied nor Dissatisfied; 4 Somewhat Dissatisfied; 5 Very Dissatisfied].

for three questions using a polar points satisfaction scale during a web and a phone interview. In one version the number 5 was associated with “Very Satisfied”⁵; in another version the number 1 was associated with “Very Satisfied”. The authors did not find evidence that assigning the most positive category with a higher number (5 → “Very Satisfied”) shifted the distribution to the more positive side in comparison to the second version (only one of the phone comparisons was statistically significant).

When graphical and symbolic languages that help communicate the continuum of the scale are removed, the number-label association becomes more important. In a paper and pencil questionnaire administered to WSU students, Christian and Dillman (2004) asked respondents to answer some five point satisfaction scales by entering a number in an answer box. For example: “On a scale of 1 to 5, with 1 being very satisfied and 5 being very dissatisfied, how satisfied are you with the classes you are taking this semester? □ Number of your rating”. The authors found that 10% of respondents scratched out answers to at least one of the questions and provided a different answer. Most of these errors occurred due to respondents reversing the scale on the answer box.

We found two studies where the number-label association was manipulated for a fully labeled ordinal scale, similar to the one used in our study. In an experiment conducted on the CenterPanel, a probability based online panel based in the Netherlands, Toepel, Das and Van Soest (2006) manipulated the number-label association for two unipolar questions with a fully labeled answer scale presented in vertical format. In one question⁶ the numbers ranged from either 1 to 5 or to 5 to 1. In another question⁷ the numbers ranged from 1 to 5 or from 2 to -2. When comparing the number-label association when the numbers ranged from 1 to 5 vs. from 5 to 1, the authors found no statistically significant difference. When comparing the 1 (Excellent) to 5 (Poor) scale with 2 to -2 the authors found a shift to the scale on the positive side. This shift did not reach a significant statistical significance in the chi square test. In a more recent study (Toepel, Das & van Soest, 2008), the same authors redid the previous experiment, contrasting the same answer scale⁸ with numbers going from 1 to 5 (group a), 5 to 1 (group b), and -2 to +2 (group c) in a sample of fresh or trained respondents. This time the results were different; when comparing group 1 to 2, the authors noticed a drop in the number of people endorsing very good, when it was associated with the number 4 (group b) instead of 2 (group a). The label “Very good” was again the most affected item when comparing group b (number 4 associated with) to group c (number 1 associated with). The percent of people endorsing “very good” increased from 9% to 19.9% for the trained respondents and from 7.6 to 13.3% for the fresh respondents.

Much more literature on response alternatives order effect has been published than on the previous topic. The majority of research has been done on categorical or unordered response alternatives order effect [for a summary see Schuman & Presser chapter 2 (1981); Sudman, Bradburn & Schwarz chapter 6 (1996); and Krosnick, Judd & Wittenbrink (2005)]. The general finding is that order effect does matter, and it is linked to the mode of presentation of the question. For visually presented items primacy effects are more prominent, while for auditorily presented items recency effects are more

⁵ [5 Very Satisfied;4;3;2;1 Very Dissatisfied] [1 Very Satisfied; 2; 3; 4; 5 Very Dissatisfied].

⁶ Overall, how would you rate the quality of education in the Netherlands? [Excellent, Very good, Good, Fair, Poor]

⁷ Overall, how would you rate the quality of life in the Netherlands? [Excellent, Very good, Good, Fair, Poor]

⁸ Overall, how would you rate the quality of education in the Netherlands? [Excellent, Very good, Good, Fair, Poor]

likely to manifest. When concentrating on order of presentation for rating scales (the focus of this paper), the story is different. From the experiments conducted, it appears that primacy effects are present for both visual and aural presentation (Krosnick et al., 2005; Sudman et al., 1996, p. 157–8).

There is mounting evidence that numbers used to label rating scale points often selected arbitrarily for the goal of precoding (Krosnick, 1999), do have an effect on the response distribution. The effect is present no matter the mode of administration: visual and auditory (Schaeffer & Baker, 1995; Schwarz & Hippler, 1995). In fact, information is signaled to the respondent in the question stem (O'Muircheartaigh et al., 1995), and it depends on the choice of numbers used as endpoints (Schwarz et al., 1998). Negative numbers seem to shift the distribution even more to the positive side than when not used (O'Muircheartaigh et al., 1995). Schaeffer and Presser (2003) and Schwarz (2008) remind us then, when negative numbers are used, they transform the scale from unipolar to bipolar, thus changing the measurement properties of the scale (O'Muircheartaigh et al., 1995). Moreover the positivity bias theory suggests that respondents are reluctant to give negative evaluations. This is more prevalent where negative numbers are used to anchor the scale (Tourangeau, Rips, & Rasinski, 2000, p. 241 and seq.). Respondents do use all the information given by the question and response alternatives, plus its visual layout, to disambiguate its meaning, thus including the numbers in the interpretation. When the visual layout and other graphics features are absent, the number-label association becomes even more prominent (Christian & Dillman, 2004).

Response order effects for categorical questions are explained by weak satisficing (Krosnick, 1999). Respondents applying this behavior choose the first response option that is reasonable instead of carefully considering the entire list and then making their judgment. The weak satisficing notion explains primacy effects in self administered questionnaires and recency effects in auditorily presented questionnaires. What is more difficult to explain is why, when considering response order effects for rating scales, the tendency is to produce primacy effects independently from the mode of data collection [but see Dillman et al. (1995) for contrasting results]. Sudman and colleagues (1996, p. 157-8) explain this phenomenon with the fact that a rating scale does not need to be fully elaborated, because the respondents can quickly infer the direction of the labels in their continuum. The authors make the hypothesis that the effect is due to anchoring of the type described by Tversky and Kahneman (1974). On the other hand, Krosnick et al. (2005) explain it with the weak satisficing paradigm: "If a satisficing participant considers the options on a rating scale sequentially, then he or she may select the first one that falls in his or her latitude or acceptance, yielding a primacy effect under both visual and oral presentation" (p. 45). Both groups of researchers conclude advocating for more research on the topic.

In the present study we wanted to explore the relationship between numbers and verbal label association for a fully labeled five point rating scale⁹. We selected a five point bipolar satisfaction scale, because it is very common in questionnaires. In this

⁹ In the midst of writing this paper we became aware of a paper that did a similar experiment to our study (Meric & Wagner, 2006). The authors used a six point agree-disagree scale with number association on a paper and pencil survey with undergraduate students. We however decided not to report their findings because we noted a confounding element in the study. The spacing of the columns where the response options were reported (grid) was completely uneven (e.g. "somewhat agree" took almost double the space of the "very strongly agree" column). Uneven spacing does change the response distribution as shown by Tourangeau, Couper and Conrad (2004) and therefore the results of the Meric & Wagner study are questionable.

experiment we decided to use numbers going from 1 to 5. The main goal was to measure what happens when the numbers communicate conflicting meanings with the labels, for example 5 associated with “very unsatisfied”. As Krosnick (1999) suggests: “... rating scale points should be labeled only with words or [with] numbers [that] should reinforce the meaning of the words” (p. 544). Our hypothesis was that when respondents would find conflicting meaning between numbers and labels, they would spend more time on those items. In order to control for other confounding factors, we also manipulated the vertical order of presentation of the response alternatives. The second hypothesis was to encounter and measure the extent of primacy effects for rating scales.

METHOD

A sample of Knowledge Panel[®] participants was selected for the present study. Knowledge Networks initially selects households using random digit dialing (RDD) sampling methodology. Once a household is contacted by phone and household members recruited to the panel by obtaining their e-mail address or setting up e-mail addresses, panel members are sent surveys invitations over the internet using e-mail. When the respondents click to the unique survey link, they are redirected to our servers, where they are administered a web survey. As of August 2002, those RDD households that inform interviewers that they have a home computer and internet access have been recruited to the panel. Knowledge Networks asks these households to take surveys using their own equipment and internet connections. If the household does not have a PC and access to the internet, they are told that in return for completing a short survey weekly, the household will be given a WebTV set-top box, or in some cases a PC, and free monthly internet access.

STUDY 1: MANIPULATION OF NUMBER LABEL ASSOCIATION FOR A FULLY LABELED BIPOLAR SCALE

The survey was fielded during the month of April 2008 and 3,667 participants completed the study for a 71% completion rate. The break-off rate was of 1.7%. The study had a 2 x 2 experimental design plus two control groups. The first factor (groups 1 and 2) was to vary the number associated (1 to 5 vs. 5 to 1) to a satisfaction scale going from very satisfied to very unsatisfied. In the second factor (groups 3 and 4), we reversed the order of presentation of the first two versions of the questionnaire (e.g., starting from very unsatisfied) keeping the number-label association constant. In the control groups (groups 5 and 6) no number was associated with the satisfaction scale, but in group 6 we reversed the order of the scale. Approximately 610 people completed each condition. Table 1 shows the six experimental conditions, while Appendix A reports the actual screenshots used in the survey.

Table 1. Experimental Design for Study 1

<p>Group 1</p> <p>1 Very Satisfied 2 Somewhat Satisfied 3 Neither Satisfied nor Unsatisfied 4 Somewhat Unsatisfied 5 Very Unsatisfied</p>	<p>Group 2</p> <p>5 Very Satisfied 4 Somewhat Satisfied 3 Neither Satisfied nor Unsatisfied 2 Somewhat Unsatisfied 1 Very Unsatisfied</p>
<p>Group 3</p> <p>1 Very Unsatisfied 2 Somewhat Unsatisfied 3 Neither Unsatisfied nor Satisfied 4 Somewhat Satisfied 5 Very Satisfied</p>	<p>Group 4</p> <p>5 Very Unsatisfied 4 Somewhat Unsatisfied 3 Neither Unsatisfied nor Satisfied 2 Somewhat Satisfied 1 Very Satisfied</p>
<p>Group 5</p> <p>Very Satisfied Somewhat Satisfied Neither Satisfied nor Unsatisfied Somewhat Unsatisfied Very Unsatisfied</p>	<p>Group 6</p> <p>Very Unsatisfied Somewhat Unsatisfied Neither Unsatisfied nor Satisfied Somewhat Satisfied Very Satisfied</p>

The study consisted of seven attitudinal questions on satisfaction about some aspects of life in the U.S. The exact question wording is presented in Appendix A. Each item was presented in a single screen, and the order of presentation of the questions was randomized within each condition. The survey is programmed so WebTV participants see the same exact screen as the PC participants, no extra scrolling is required.

RESULTS

Chi square test analysis comparing level of education, age, race, gender and WebTV versus PC ownership revealed that the random allocation to the six groups really worked. Therefore in no group were these demographics and PC ownership variables overrepresented.

The first analysis was run to verify if adding numbers to the scale did change the distribution of the answers, in contrast to the scales with no numbers. The comparison was done among the groups where the order of presentation of the scale was the same as shown in Table 2. For the analysis we used chi-square tests, treating the response options as ordinal.

Table 2. Chi-square Tests for Effect of Number Presence, Keeping Order Constant

	Group 1 vs. 5	Group 2 vs. 5	Group 3 vs. 6	Group 4 vs. 6
Question a	2.43	5.59	3.89	3.65
Question b	16.03*	0.41	2.64	3.19
Question c	3.13	4.82	10.28*	3.83
Question d	2.00	8.63	4.64	4.91
Question e	7.10	4.48	3.47	2.62
Question f	11.37*	7.10	1.69	5.57
Question g	1.77	3.77	3.86	8.35

P<.05

4df, cv=9.49

From Table 2 we can infer that the presence of numbers for a fully labeled bipolar scale does not make any difference in the distribution of the variables.

The following analysis was done to understand if reversing the number-label associations would change the distribution of the data, by keeping the order of presentation constant.

Table 3. Chi-Square Tests for Effect of Number Reversal, Keeping Label Order Constant

	Group 1 vs. 2	Group 3 vs. 4
Question a	1.23	5.92
Question b	14.83*	9.59*
Question c	12.12*	3.71
Question d	9.78*	6.58
Question e	4.86	2.22
Question f	2.36	11.63*
Question g	5.17	4.11

From Table 3 we can infer that reversing the numbers associated with the response options do make some difference, although the pattern is not very clear. Further analyses comparing the percentage of people selection¹⁰ each response option with the respective response options in the other experimental condition gave extra information. When presented with a scale numbered 5 to 1 (rather than 1 to 5), substantially more respondents selected the middle category (“neither satisfied nor dissatisfied”). An example is shown in Appendix C1.

¹⁰ Z test of proportions

In the next comparison we tested the effect of reversing the scale, keeping the number association constant. We also contrasted the control groups (5 and 6).

Table 4. Chi-Square Tests for Effect of Label Order Reversal

	Group 1 vs. 4	Group 2 vs. 3	Group 5 vs. 6
Question a	8.64	5.89	7.62
Question b	10.09*	18.99*	10.14*
Question c	12.40*	23.67*	6.96
Question d	5.70	16.27*	6.08
Question e	2.22	12.16*	10.64*
Question f	1.32	17.28*	9.53*
Question g	19.64*	8.66	14.86*

P<.05

4df, cv=9.49

The reversing of the scale did create an effect on the distribution of the answers. In the first two comparisons the numbers are not reversed, but they stay associated with the label (e.g. 1 very satisfied). In the last comparison (group 5 vs. 6) there are no numbers and the scale is just reversed. The pattern of the data suggests a presence of primacy effect. When comparing the percentages of answers for each category, we found that when “very satisfied” was presented first, it was significantly endorsed more times than when presented last. Conversely, substantially more people select “very unsatisfied” when presented first. The effect is stronger for the comparison of group 2 vs. group 3. In this case the meaning of the numbers matches the labels (5 very Satisfied). An example is shown in Appendix C2.

In the last analysis, group 1 vs. 3 and group 2 vs. 4 were contrasted. Here, both the numbers and the order of the scale are reversed. Interestingly enough we do not find evidence for primacy effects when the scale is reversed.

Table 5. Chi-Square Test for Effects of Simultaneous Label Order and Number Order Reversal

	Group 1 vs. 3	Group 2 vs. 4
Question a	6.16	4.62
Question b	6.18	3.18
Question c	13.80*	17.45*
Question d	2.21	7.55
Question e	6.22	6.59
Question f	9.36	4.83
Question g	8.65	13.48*

P<.05

4df, cv=9.49

Time Latency Analysis

In the initial hypothesis we put forward the idea that were the number-label association to be reversed (e.g., 5 for “very satisfied”), it would confuse people, and therefore the respondent had to take some extra time to figure out the scale. The analysis using a Mann Whitney median test partially supported the hypothesis. Time latency for group 1 vs. group 2 did not reach a statistically significant difference, while it did when comparing group 3 to group 4. In the latter case people in group 4 take one second more to answer the first question shown (median of 13 vs. 14 seconds). The action is concentrated in the question shown first, after that the respondents take more or less the same time to reply.

The main trend in the analysis is that respondents in group 3 and 4 take more time than respondents in group 1 and 2, again mostly for the first question shown. We explain the findings by the fact that in group 3 and 4, the scale is reversed and does not flow well in the question wording. In the question stem we ask how satisfied people are, and the first response option shown is “very unsatisfied”. The finding is consistent with or without the number-labels associations, e.g., group 5 is statistically different from group 6—the latter taking longer to reply to the first question shown.

STUDY 2: MANIPULATION OF NUMBER LABEL ASSOCIATION FOR AN ENDPOINT LABELED SCALE

The survey was fielded during the month of August 2008, and 3,273 participants completed the study for a 72% completion rate. The break-off rate was of 0.9%. The study had a 2 x 2 experimental design. The first factor (groups 1 and 2) was to vary the number associated (1 to 5 vs. 5 to 1) to a polar point labeled satisfaction scale going from “very satisfied” to “very unsatisfied”. In the second factor (groups 3 and 4) we reversed the order of presentation of the first two versions of the questionnaire (e.g. starting from “very unsatisfied”) keeping the number-label association constant. Table 6 shows the six experimental conditions, while Appendix D reports the actual screenshots used in the survey.

Table 6. Experimental Design for Study 2

Group 1 1 Very Satisfied 2 3 4 5 Very Unsatisfied	Group 2 5 Very Satisfied 4 3 2 1 Very Unsatisfied
Group 3 1 Very Unsatisfied 2 3 4 5 Very Satisfied	Group 4 5 Very Unsatisfied 4 3 2 1 Very Satisfied

The study consisted in seven attitudinal questions about the state of the nation in several different areas. The exact question wording is presented in Appendix E. The questions were different than the questions in study 1. Each item was presented in a single screen, and the order of presentation of the questions was randomized within each condition. The survey is programmed so WebTV participants see the same exact screen as the PC participants, no extra scrolling is required.

RESULTS

Chi square test analysis comparing level of education, age, race, gender and WebTV versus PC ownership revealed that the random allocation to the six groups really worked. Therefore in no group were these demographics and PC ownership variables overrepresented.

In the first analysis we compared the effect of reversing the numbers by keeping the order constant. Three of the seven comparisons between group 3 and group 4 were statistically significant. For these three items the distribution for group 4 was slightly shifted to the bottom of the scale ("Very satisfied") in comparison to group 3.

Table 7. Chi-Square Tests for Effect of Number Reversal by Keeping Label Order Constant

	Group 1 vs. 2	Group 3 vs. 4
Question a	1.76	9.55*
Question b	3.26	9.39
Question c	1.37	4.09
Question d	6.99	14.27*
Question e	4.84	1.24
Question f	4.43	10.94*
Question g	.51	4.06

P<.05

4df, cv=9.49

Z-test of proportions comparing the bottom two boxes (“Very satisfied” and its next point) showed a statistically significant difference between group 3 and 4 for question a and d. For question f the difference reached statistical significance only when comparing the polar point (“Very satisfied”) but still in the same direction as before.

In the second analysis we tested the effect of reversing the order of the labels by keeping the number-label association the same. None of the comparisons reached statistical significance.

Table 8. Chi-Square Tests for Effect of Label Order Reversal

	Group 1 vs. 4	Group 2 vs. 3
Question a	4.39	8.08
Question b	1.96	3.33
Question c	1.68	2.45
Question d	5.81	5.19
Question e	2.24	7.28
Question f	2.86	9.19
Question g	3.31	3.01

P<.05

4df, cv=9.49

In the last comparison, we tested the effect of reversing both the order of the labels and the order of the numbers, by finding only one of the seven comparisons to be statistically significant.

Table 9. Chi-Square Test for Effects of Simultaneous Label Order and Number Order Reversal

	Group 1 vs. 3	Group 2 vs. 4
Question a	5.87	1.95
Question b	6.86	8.64
Question c	4.28	2.55
Question d	4.86	5.86
Question e	3.23	3.28
Question f	15.34*	2.58
Question g	5.43	2.39

P<.05

4df, cv=9.49

DISCUSSION

In answering survey questions respondents accomplish four tasks: comprehension; retrieval; formatting of the response; and finally, reporting (Sudman et al., 1996; Tourangeau et al., 2000). In the third stage, also called judgment and estimation, respondents try to identify the response alternative that best fits their judgment. When faced with a rating scale, the effect that seems to occur unequivocally at the formatting stage is explained by Parducci's range-frequency theory (Parducci, 1983; Parducci & Wedell, 1986). The range effect theory, applied to questionnaire design, predicts that respondents use the most extreme response options to anchor the endpoints of the rating scale. The frequency effect theory predicts that respondents tend to assign a fixed proportion of the stimuli to each response category with equal frequency. Daamen and Bie (1992) provide an example of the impact of Parducci's theory on survey results.

The work done by Dillman and colleagues (2008) show how visual design has an effect on how respondents understand the question and format the response. At the same time, the number label association is used by the respondent in interpreting the meaning of a scale (O'Muircheartaigh et al., 1995; Schwarz & Hippler, 1995; Schwarz et al., 1991) and the extremes to anchor the scale.

Lastly respondents use some heuristics in interpreting visual questionnaires (Tourangeau et al., 2004, 2007). Of interest to our paper are the "left and top means first" and "up means good". The first heuristic states that "leftmost or top items in a list of items will be seen and the "first" in some conceptual sense" (Tourangeau et al., 2004, p.371), while the second states that "with a vertically oriented list, the top item or option will be seen as the most desirable" (p. 372).

In the first study when the order of the rating scale labels is kept constant, adding the numbers appears not to make any difference in how subjects rate their judgment (Table 2). It seems that for a fully labeled scale presented vertically, the label overrides the meaning of the number. When we compared group 1 vs. 2 and group 3 vs. 4 (Table 3), we found some evidence that the numbers had an effect on the scale, although not very strong. In groups 1 and 4 the meaning of the number was going in the opposite direction than the meaning of the label (e.g., 5 is "Very Unsatisfied"). But again, because the scale was fully labeled it is possible that the respondents paid more attention to the label than the number.

Primacy effects were found when the scale was reversed for each of the combinations (Table 4), with the strongest effect when the number-label association matched. The primacy effect also reproduced previous findings for rating scales (Krosnick et al., 2005). Categories at the top of the list have more chances to be endorsed than categories at the bottom of the list. A similar effect is also found for rating scales presented in horizontal order, when items on the left part of the scale have more chances to be selected than items on the right of the scale (Friedman, Herskovitz, & Pollack, 1993). The results of Table 5 are intriguing. It seems that the number-label association is counteracting the primacy effects we found in Table 4, but more research is needed on this subject.

Lastly, reversing the scale perplexes the respondents at first, and they take an extra second to answer the first question. This is because it goes against the two interpretative heuristics previously described. As the authors were expecting: "if the list does not conform to these expectations, respondents might become confused, make mistakes and take longer to respond" (Tourangeau et al., 2004, p. 371). Reversing the number-label association in meaning (lower number associated with very satisfied) still seems to confuse the respondents, although we do not have a clear picture on that.

Results from the second study suggest that for a polar point labeled scale the number-label association does not make much of a difference, although we found some effects when reversing the numbers in group 3 vs. group 4. The reader should note that we used numbers from 1 to 5 and not negatives numbers or 0.

CONCLUSIONS AND INITIAL RECOMMENDATIONS

When a rating scale is fully labeled, numbers play less of a role in aiding the respondent in interpreting the labels. Respondents were slightly confounded when the meaning of the numbers was going in the opposite direction than the meaning of the label.

Primacy effects for rating scales are of concern for survey researchers. If for unordered answer scales the common suggestion is to randomize the order of presentation, to give each item the same chance to be first on the list (Sudman et al., 1996)¹¹, the issue of randomizing a rating scale is very controversial. A rating scale is by definition conveying an order of response options; randomizing the labels probably would confuse the respondent. Two studies actually show how randomizing the order of a rating scale

¹¹ The authors remind the reader that randomization eliminates order effects in a rather mechanical way (p. 162).

does dramatically change the distribution of the answers (Friedman et al., 1993), and also decrease the number of respondents picking the middle point of the scale (Garland & Krosnick, 2007). We reserve recommendations on this topic for future work. The results from our Table 5 seem to suggest that we can take care of primacy effects by reversing the scale and reversing the numbers association with the labels. It is, however, premature to recommend using this strategy, and we want to have more evidence before taking a stand on this topic.

When a scale is fully labeled we do recommend not using any number, because it introduces an extra element of judgment which seems unnecessary, although using a -2 to +2 would make the scale clearly bipolar. This strategy would follow Krosnick's recommendation to use numbers that reinforce the meaning of the words, but at the same time it can be confounded by the positivity bias (Tourangeau et al., 2000). In only one study, and for one item, groups answering a unipolar scale question and having the number label associations going from a 1 to 5 or a 2 to -2 did not answer differently (Toepoel et al., 2006). This sparse evidence cannot rule out possible positivity bias, therefore more research is needed.

The problem occurs when researchers use only a polar point labeled scale (e.g., in a grid) mostly for practical purposes. In that case a fully labeled scale would make the width of the columns so large that the entire table would not fit in the screen. In these cases (grid with polar point labels only), the researchers should expect a shift in the distribution of answers, compared to a fully labeled scale as demonstrated by Dillman & Christian (2005), although the authors did not experiment with grids but with one question per screen only.

It seems also advisable to start the scale in the same order given by the question stem. If the question stem starts with "How satisfied...", the respondents expect to find a "satisfied" response option first and not the contrary. Our time latency analysis supported this hypothesis. This advice is in line with the interpretative heuristics delineated by Tourangeau, Couper and Conrad (2004).

Acknowledgements

Erica Demme and Patricia Graham provided useful comments on the text increasing readability and clarity of the exposition. Jing Yan Quek provided assistance in deploying the survey and doing quality control. Brian Spiegel scripted study number 1, while Frank Chang study number 2.

REFERENCES

- Abrams, J. (1966). An evaluation of alternative rating devices for consumer research. *Journal of Marketing Research*, 3, 189–183.
- Bourque, L. B., & Fielder, E. P. (1995). *How to conduct self-administered and mail surveys*. Thousand Oaks, CA: Sage.
- Christian, L. M. (2003). *The influence of visual layout on scalar questions in Web surveys*. Unpublished master's thesis. Pullman, WA: Washington State University.
- Christian, L. M., & Dillman, D. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public Opinion Quarterly*, 68, 57–80.
- Christian, L. M., Dillman, D. A., & Smyth, J. D. (2008). The effects of mode and format on answers to scalar questions in telephone and Web surveys. In J. M. Lepkowski, C. Tucker, M. J. Brick, E. De Leeuw, L. Japac, P. J. Lavrakas, M. W. Link & R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 250–275). Hoboken, NJ: Wiley.
- Daamen, D. D. L., & de Bie, S. E. (1992). Serial context effects in survey interviews. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 97–113). New York: Springer.
- Dillman, D. A. (2007). *Mail and Internet surveys. The tailored design method* (Second ed.). Hoboken: John Wiley & Sons, Inc.
- Dillman, D. A., Brown, T. L., Carlson, J. E., Carpenter, E. H., Lorenz, F. O., Mason, R., et al. (1995). Effects of category order on answers in mail and telephone surveys. *Rural Sociology*, 60, 674–687.
- Dillman, D. A., & Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods*, 17, 30–52.
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2008). *Internet, mail and mixed-mode surveys. The tailored design method* (Third ed.). Hoboken: John Wiley & Sons, Inc.
- Friedman, H., H., Herskovitz, P. J., & Pollack, S. (1993). The biasing effect of scale-checking styles on response to a Likert scale. In A. S. Association (Ed.), *Proceedings of the Joint Statistical Meeting, Survey Research Methods section* (pp. 792–795). Alexandria, VA: AMSTAT.
- Garland, P., & Krosnick, J. A. (2007). *Response option ordering: Reconciling meanings conveyed by rating scale position and label*. Unpublished paper. Stanford University.
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50, 537–567.
- Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2005). The measurement of attitudes. In D. Albarracín, B. T. Johnson & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 21–78). Mahwah, NJ: Erlbaum.
- Meric, H., & Wagner, J. (2006). Rating scale format choices for multi-item measures: Does numbering and balanced-ness matter? *Business Quest*.
- O'Muircheartaigh, C., Gaskell, G., & Wright, D. B. (1995). Weighting anchors: Verbal and numeric labels for response scales. *Journal of Official Statistics*, 11, 295–307.
- Parducci, A. (1983). Category ratings and the relational character of judgment. In H.-G. Geissler, H. F. J. M. Buffart, E. L. J. Leeuwenberg & V. Sarris (Eds.), *Modern issues in perception* (pp. 262–282). New York: North-Holland.

Parducci, A., & Wedell, D. H. (1986). The category effect with raring scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human perception and Performance*, 12, 496–516.

Schaeffer, N. C., & Baker, K. (1995, May). Alternative methods of presenting bi-polar scales in telephone interviews: 1 to 7 vs. -3 to +3 and neutral vs. ambivalent. Paper presented at the 50th Annual meeting of the American Association for Public Opinion Research, Fort Lauderdale, FL.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, 29, 65–88.

Schuman, H., & Presser, S. (1981). *Question & answers in attitude surveys. Experiment of question form, wording, and context.* San Diego, CA: Academy Press.

Schwarz, N. (2008). The psychology of survey response. In W. Donsbach & M. W. Traugott (Eds.), *The Sage handbook of public opinion research* (pp. 374–387). London: Sage.

Schwarz, N., Grayson, C. E., & Knäuper, B. (1998). Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research*, 10, 177–183.

Schwarz, N., & Hippler, H.-J. (1995). The numeric values of rating scales: A comparison of their impact in mail surveys and telephone interviews. *International Journal of Public Opinion Research*, 7, 72–74.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, 55, 570–582.

Smyth, J. D., Dillman, D., & Christian, L. M. (2007). Context effects in internet surveys. *New issues and evidence.* In A. N. Joinson, K. Y. A. McKenna, T. Postmes & U.-D. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 429–445). oxford: Oxford University Press.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers. The application of cognitive processes to survey methodology.* San Francisco: Jossey Bass.

Toepoel, V., Das, M., & van Soest, A. (2006). *Design of web questionnaires: The effect of layout in rating scales.* Unpublished working paper. CentERdata Tilburg University

Toepoel, V., Das, M., & van Soest, A. (2008). Effects of design in web surveys: Comparing trained and fresh respondents *Public Opinion Quarterly*, 72.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, position, and order interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, 68, 368–393.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly*, 71, 91–112.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.

APPENDIX A

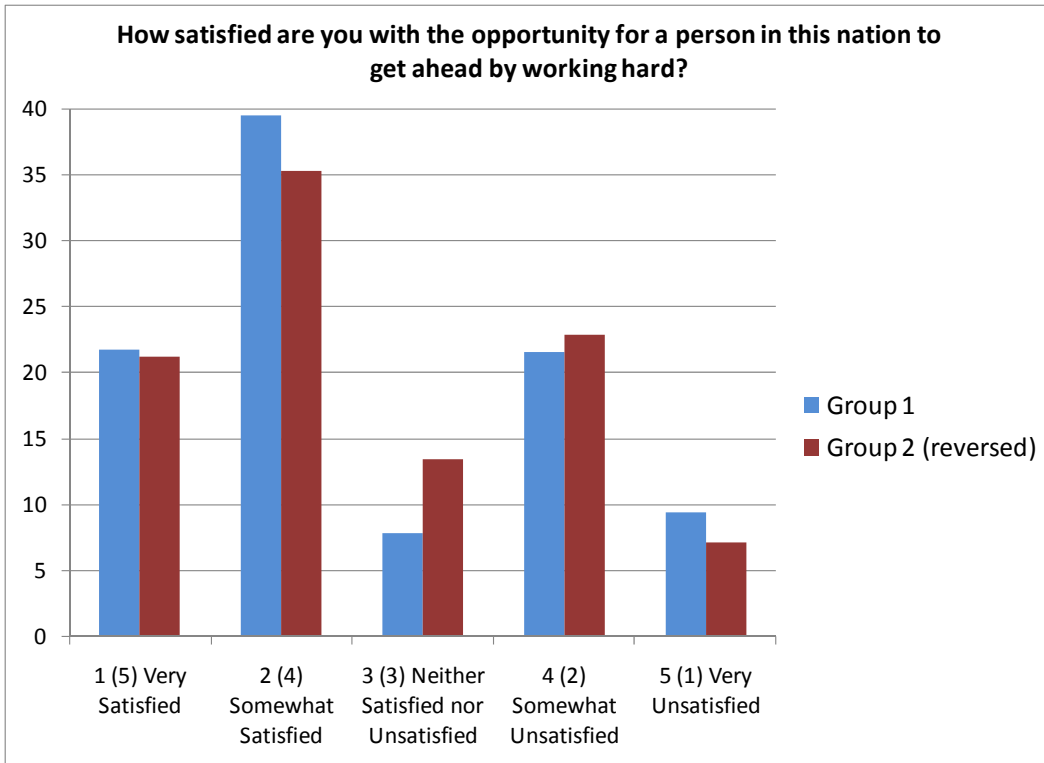
Group 1	Group 2
<p data-bbox="272 289 748 310">How satisfied are you with the overall quality of life?</p> <p data-bbox="272 344 423 361">Select one answer only</p> <ul data-bbox="313 401 586 520" style="list-style-type: none"><input type="radio"/> 1 Very Satisfied<input type="radio"/> 2 Somewhat Satisfied<input type="radio"/> 3 Neither Satisfied nor Unsatisfied<input type="radio"/> 4 Somewhat Unsatisfied<input type="radio"/> 5 Very Unsatisfied <p data-bbox="781 659 841 676">Next</p>	<p data-bbox="899 289 1375 310">How satisfied are you with the overall quality of life?</p> <p data-bbox="899 344 1050 361">Select one answer only</p> <ul data-bbox="940 401 1213 520" style="list-style-type: none"><input type="radio"/> 5 Very Satisfied<input type="radio"/> 4 Somewhat Satisfied<input type="radio"/> 3 Neither Satisfied nor Unsatisfied<input type="radio"/> 2 Somewhat Unsatisfied<input type="radio"/> 1 Very Unsatisfied <p data-bbox="1411 651 1471 667">Next</p>
Group 3	Group 4
<p data-bbox="272 745 748 766">How satisfied are you with the overall quality of life?</p> <p data-bbox="272 800 423 816">Select one answer only</p> <ul data-bbox="313 856 586 976" style="list-style-type: none"><input type="radio"/> 1 Very Unsatisfied<input type="radio"/> 2 Somewhat Unsatisfied<input type="radio"/> 3 Neither Unsatisfied nor Satisfied<input type="radio"/> 4 Somewhat Satisfied<input type="radio"/> 5 Very Satisfied <p data-bbox="781 1115 841 1131">Next</p>	<p data-bbox="899 745 1375 766">How satisfied are you with the overall quality of life?</p> <p data-bbox="899 800 1050 816">Select one answer only</p> <ul data-bbox="940 856 1213 976" style="list-style-type: none"><input type="radio"/> 5 Very Unsatisfied<input type="radio"/> 4 Somewhat Unsatisfied<input type="radio"/> 3 Neither Unsatisfied nor Satisfied<input type="radio"/> 2 Somewhat Satisfied<input type="radio"/> 1 Very Satisfied <p data-bbox="1411 1104 1471 1121">Next</p>
Group 5	Group 6
<p data-bbox="272 1201 748 1222">How satisfied are you with the overall quality of life?</p> <p data-bbox="272 1255 423 1272">Select one answer only</p> <ul data-bbox="313 1312 570 1432" style="list-style-type: none"><input type="radio"/> Very Satisfied<input type="radio"/> Somewhat Satisfied<input type="radio"/> Neither Satisfied nor Unsatisfied<input type="radio"/> Somewhat Unsatisfied<input type="radio"/> Very Unsatisfied <p data-bbox="781 1564 841 1581">Next</p>	<p data-bbox="899 1201 1375 1222">How satisfied are you with the overall quality of life?</p> <p data-bbox="899 1255 1050 1272">Select one answer only</p> <ul data-bbox="940 1312 1196 1432" style="list-style-type: none"><input type="radio"/> Very Unsatisfied<input type="radio"/> Somewhat Unsatisfied<input type="radio"/> Neither Unsatisfied nor Satisfied<input type="radio"/> Somewhat Satisfied<input type="radio"/> Very Satisfied <p data-bbox="1411 1564 1471 1581">Next</p>

APPENDIX B

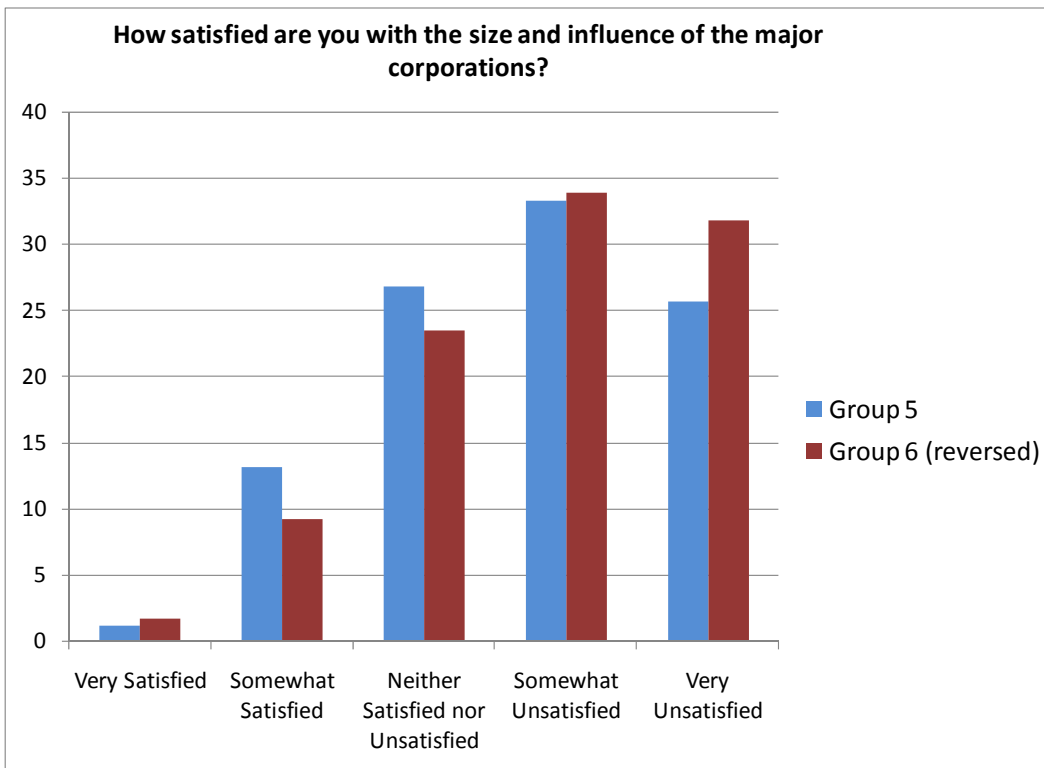
Items shown to the respondents

- a) How satisfied are you with the influence of organized religion?
- b) How satisfied are you with the size and influence of major corporations?
- c) How satisfied are you with the moral and ethical climate?
- d) How satisfied are you with the size and power of the federal government?
- e) How satisfied are you with our system of government and how well it works?
- f) How satisfied are you with the overall quality of life?
- g) How satisfied are you with the opportunity for a person in this nation to get ahead by working hard?

APPENDIX C1



APPENDIX C2



APPENDIX D

Group 1	Group 2
<p data-bbox="256 277 781 323">How satisfied are you with the nation's campaign finance laws?</p> <p data-bbox="256 352 412 373">Select one answer only</p> <ul data-bbox="298 407 456 533" style="list-style-type: none"><input type="radio"/> 1 Very Satisfied<input type="radio"/> 2<input type="radio"/> 3<input type="radio"/> 4<input type="radio"/> 5 Very Unsatisfied <p data-bbox="764 646 813 667">Next</p>	<p data-bbox="883 277 1408 323">How satisfied are you with the nation's campaign finance laws?</p> <p data-bbox="883 352 1039 373">Select one answer only</p> <ul data-bbox="925 407 1083 533" style="list-style-type: none"><input type="radio"/> 5 Very Satisfied<input type="radio"/> 4<input type="radio"/> 3<input type="radio"/> 2<input type="radio"/> 1 Very Unsatisfied <p data-bbox="1398 646 1446 667">Next</p>
Group 3	Group 4
<p data-bbox="256 724 781 770">How satisfied are you with the nation's campaign finance laws?</p> <p data-bbox="256 800 412 821">Select one answer only</p> <ul data-bbox="298 854 456 980" style="list-style-type: none"><input type="radio"/> 1 Very Unsatisfied<input type="radio"/> 2<input type="radio"/> 3<input type="radio"/> 4<input type="radio"/> 5 Very Satisfied <p data-bbox="764 1094 813 1115">Next</p>	<p data-bbox="883 724 1408 770">How satisfied are you with the nation's campaign finance laws?</p> <p data-bbox="883 800 1039 821">Select one answer only</p> <ul data-bbox="925 854 1083 980" style="list-style-type: none"><input type="radio"/> 5 Very Unsatisfied<input type="radio"/> 4<input type="radio"/> 3<input type="radio"/> 2<input type="radio"/> 1 Very Satisfied <p data-bbox="1398 1094 1446 1115">Next</p>

APPENDIX E

Items shown to the respondents

- a) How satisfied are you with the nation's campaign finance laws?
- b) How satisfied are you with the state of the nation's economy?
- c) How satisfied are you with the quality of public education in the nation?
- d) How satisfied are you with the nation's energy policies?
- e) How satisfied are you with the availability of affordable health care?
- f) How satisfied are you with the Social Security and Medicare systems?
- g) How satisfied are you with the nation's security from terrorism?